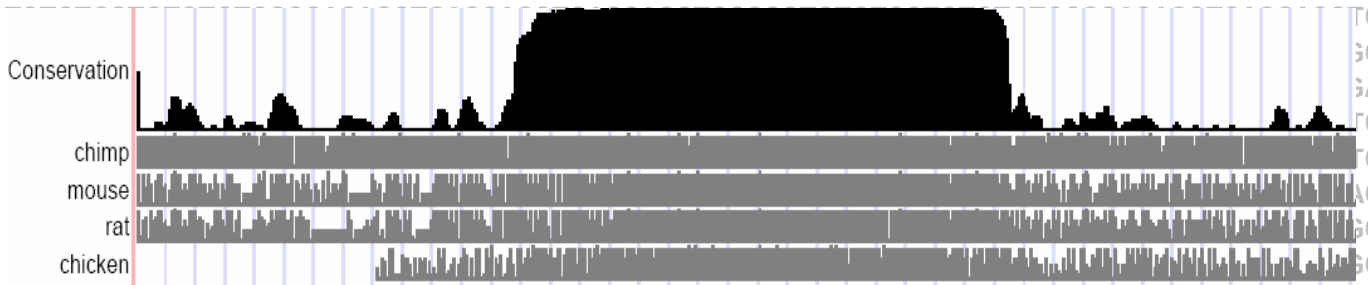
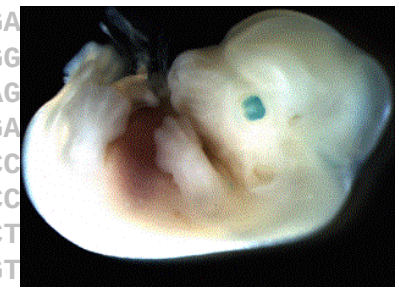
Deciphering the Human Genome: Computational Insights & Opportunities





Gill Bejerano

2006

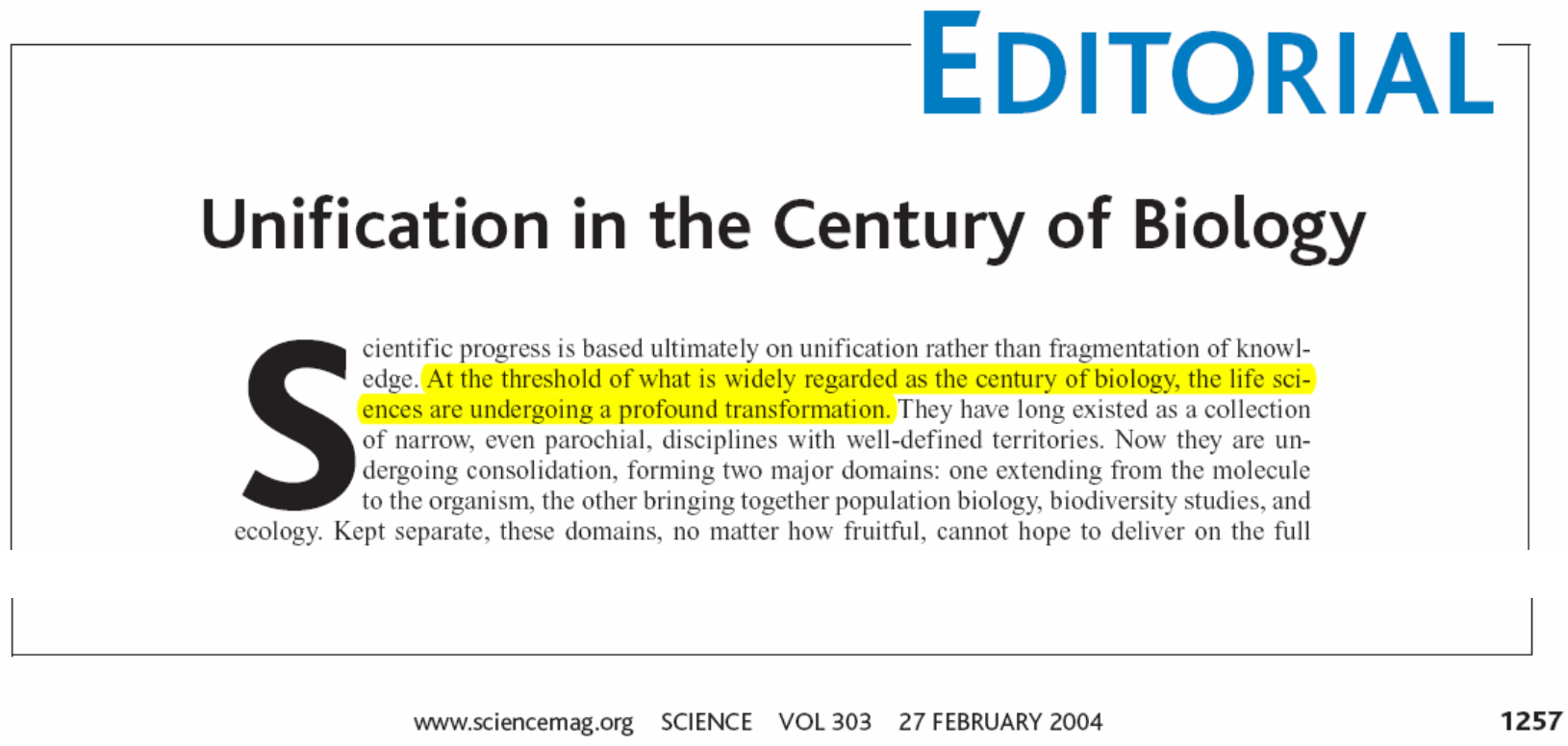
Postdoc w/David Haussler
School of Engineering
UC Santa Cruz

→

2007

Assistant Professor
Dept. of Developmental Biology
& Dept. of Computer Science
Stanford University

This is “the Century of Biology”



EDITORIAL

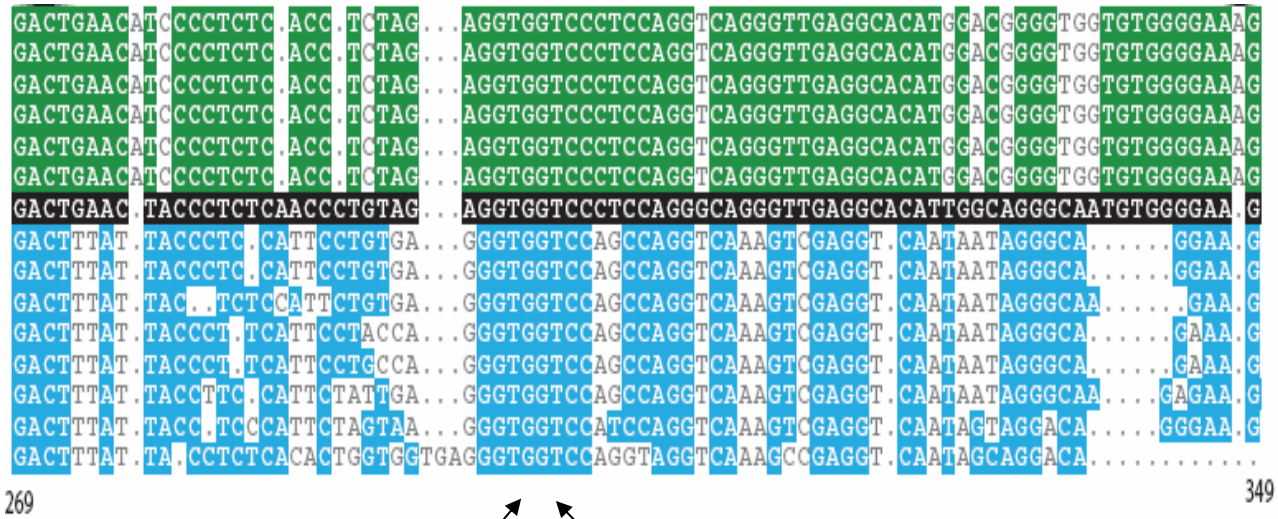
Unification in the Century of Biology

Scientific progress is based ultimately on unification rather than fragmentation of knowledge. At the threshold of what is widely regarded as the century of biology, the life sciences are undergoing a profound transformation. They have long existed as a collection of narrow, even parochial, disciplines with well-defined territories. Now they are undergoing consolidation, forming two major domains: one extending from the molecule to the organism, the other bringing together population biology, biodiversity studies, and ecology. Kept separate, these domains, no matter how fruitful, cannot hope to deliver on the full

www.sciencemag.org SCIENCE VOL 303 27 FEBRUARY 2004 1257

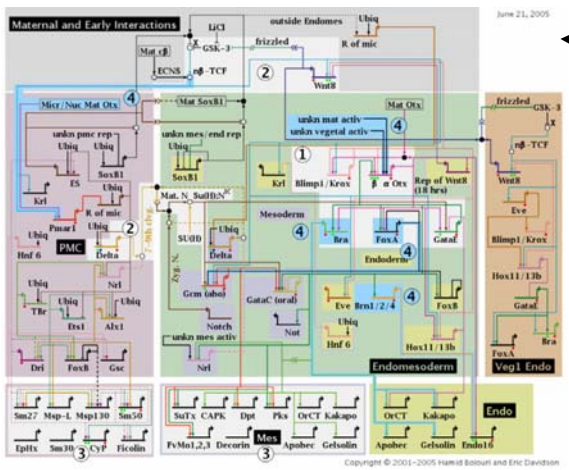
We can now cast Biology in “our” terms

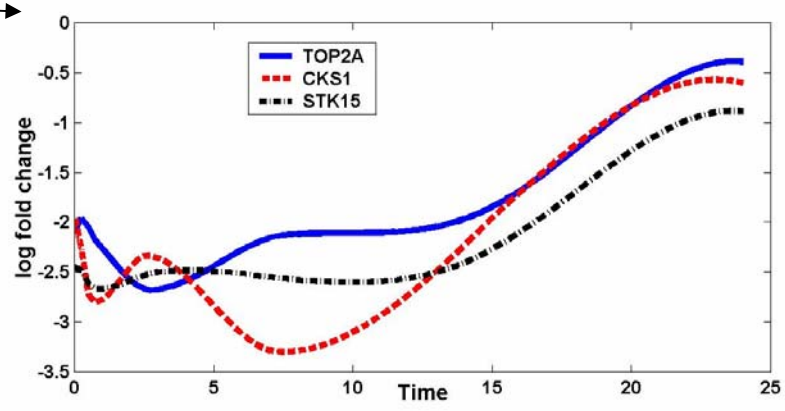
strings



time series

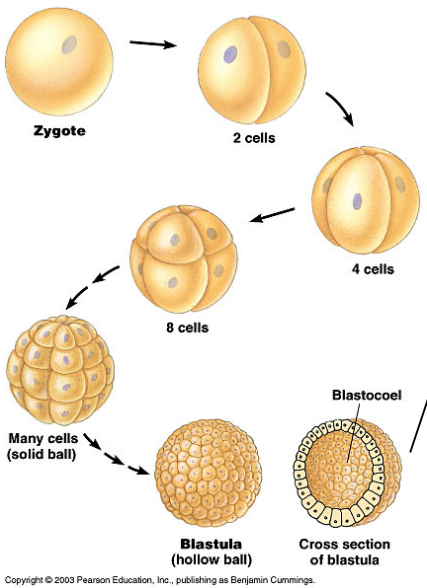
circuits

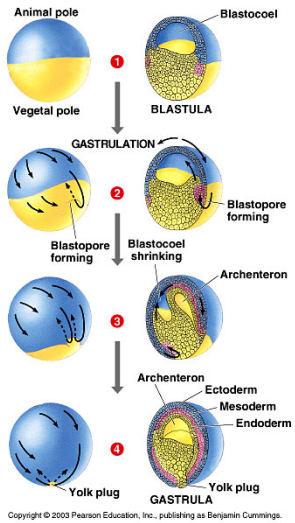




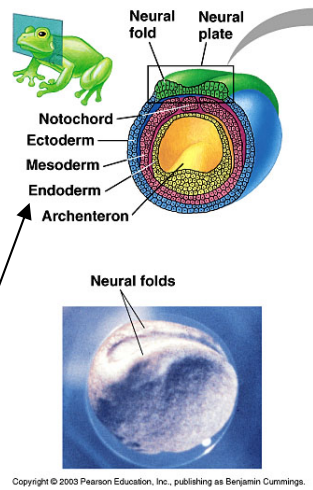
Grand Challenge: Understanding Embryonic Development

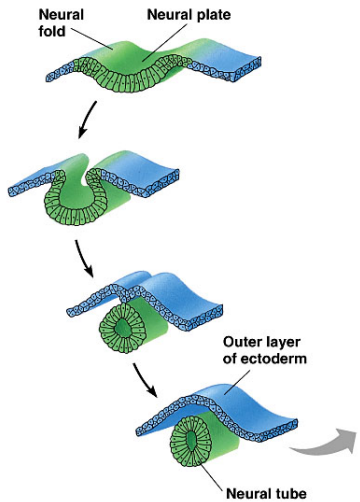
one cell





organism



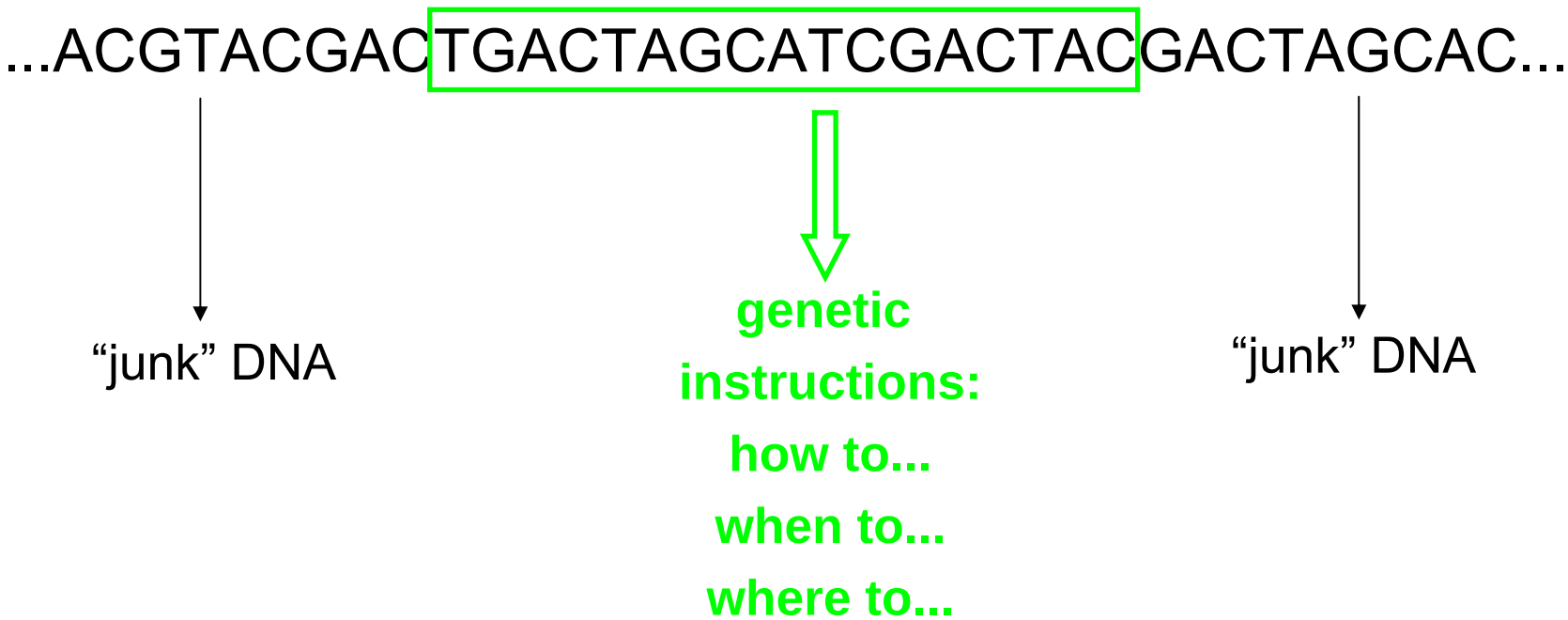


Enter DNA (“Merely the Secret of Life”) ...

DNA: Functional and Non-Functional

DNA = linear molecule that carries instructions for making living organisms ~ long string(s) over a small alphabet

Alphabet of four {A,C,G,T} Strings of length 10^4-10^{11}



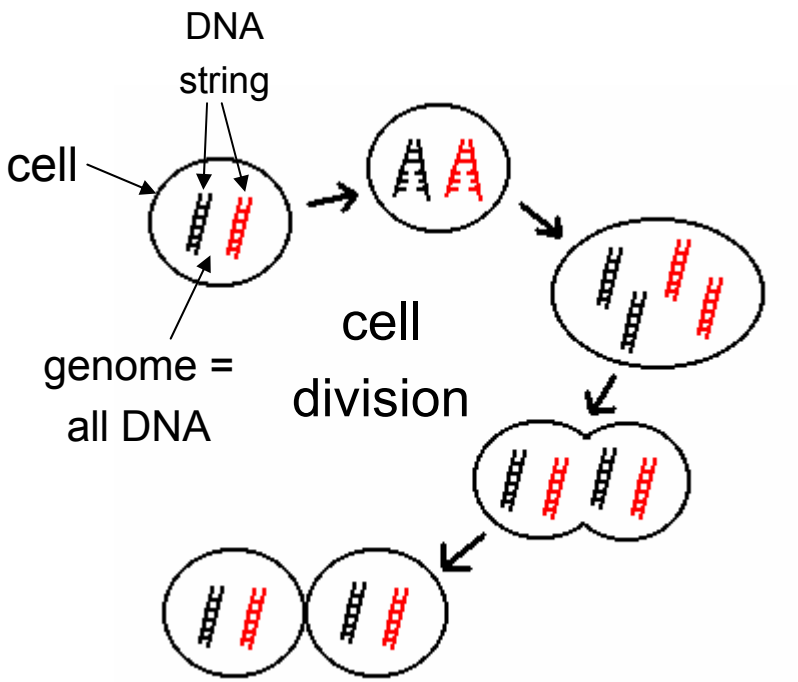
One Cell, One Genome, One Replication

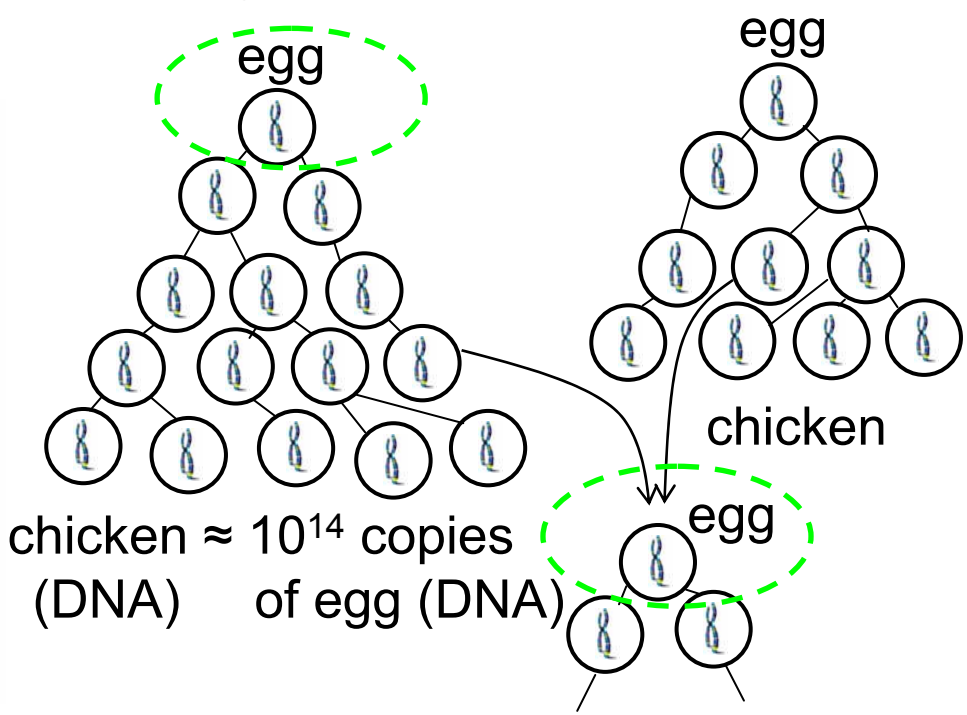
Every cell holds a single copy of all its DNA = its genome.

The genome is replicated every cell division.

The human body is made of ~10^{14} cells.

All originate from a *single* cell through cell division.



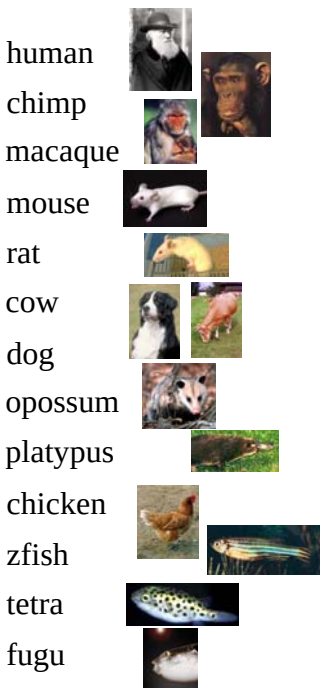


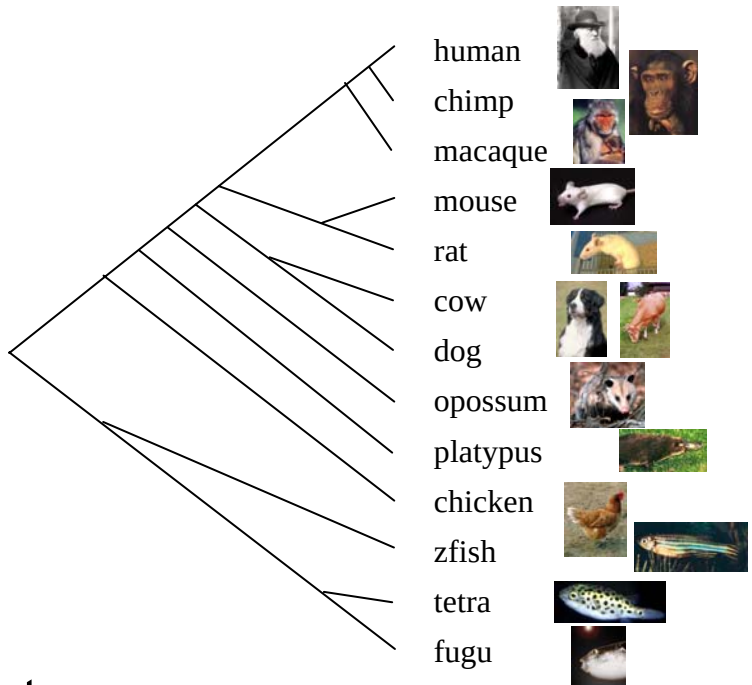
Comparative Genomics

"Nothing in Biology Makes Sense
Except in the Light of Evolution"

Theodosius Dobzhansky

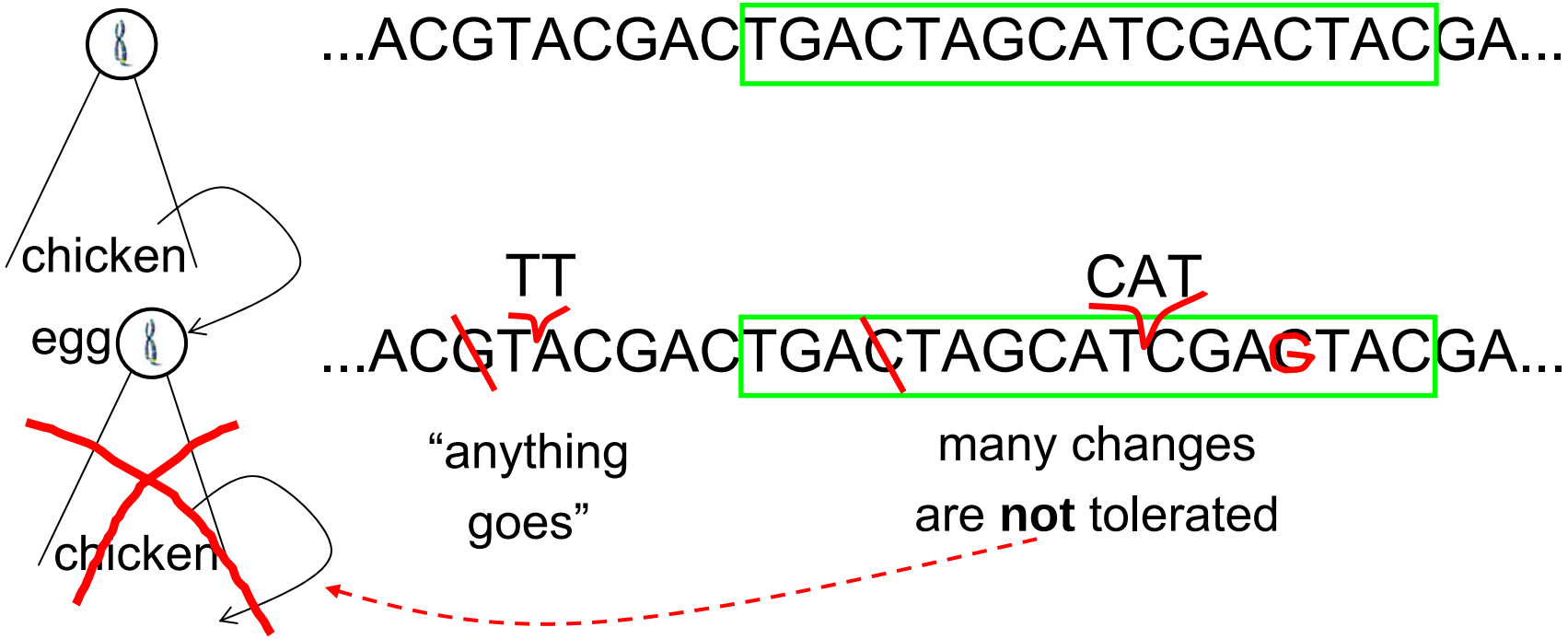
Intelligent Designer





DNA Replication is Imperfect

Small Scale: single letters are substituted, erased, added

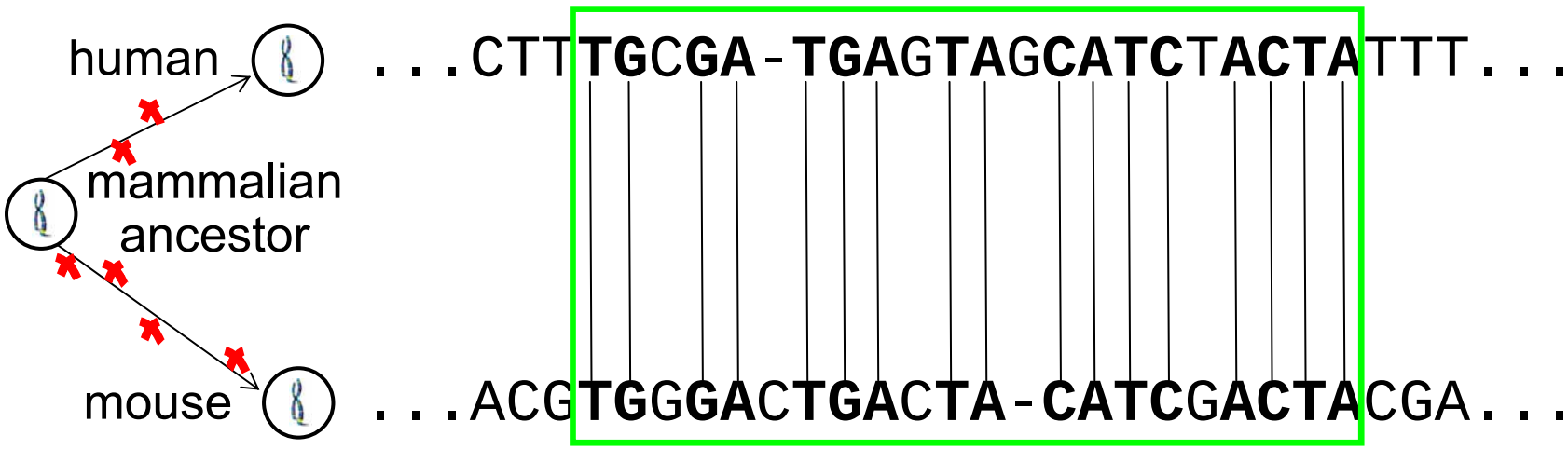


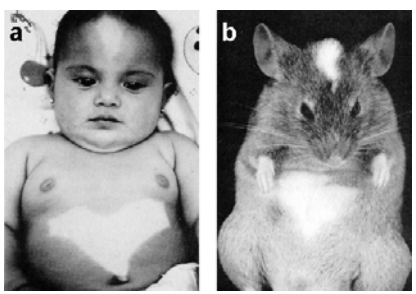
thus, sequence conservation over generations ➔ function!

Sequence Conservation implies Function

Comparative Genomics of Distantly related species:

functional region!





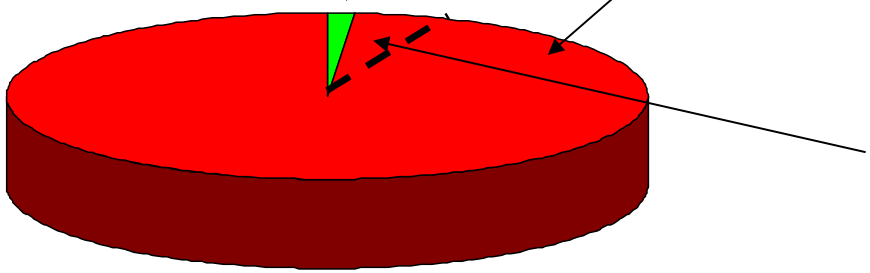
(but which function/s?...)

The Human Genome is Full of Mysteries

Human Genome: $3*10^9$ letters

1.5% known function

>50% junk



compare to other species

>5% human genome functional

3x more functional DNA than known!

But what do these ~10^7 substrings do??

[*Science* 2004 Breakthrough of the Year, 5th runner up]

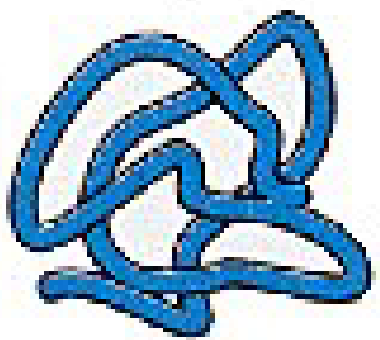
Genes = How to make Proteins

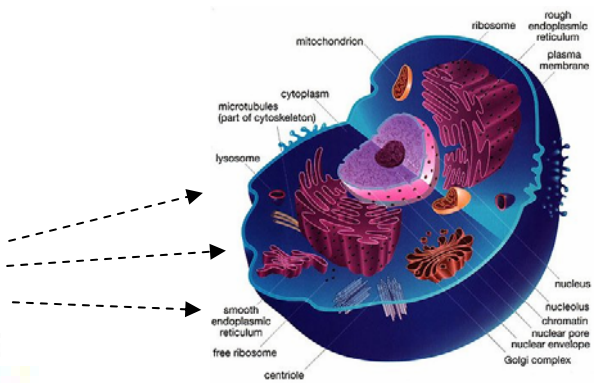
gene

DNA

Translation

protein





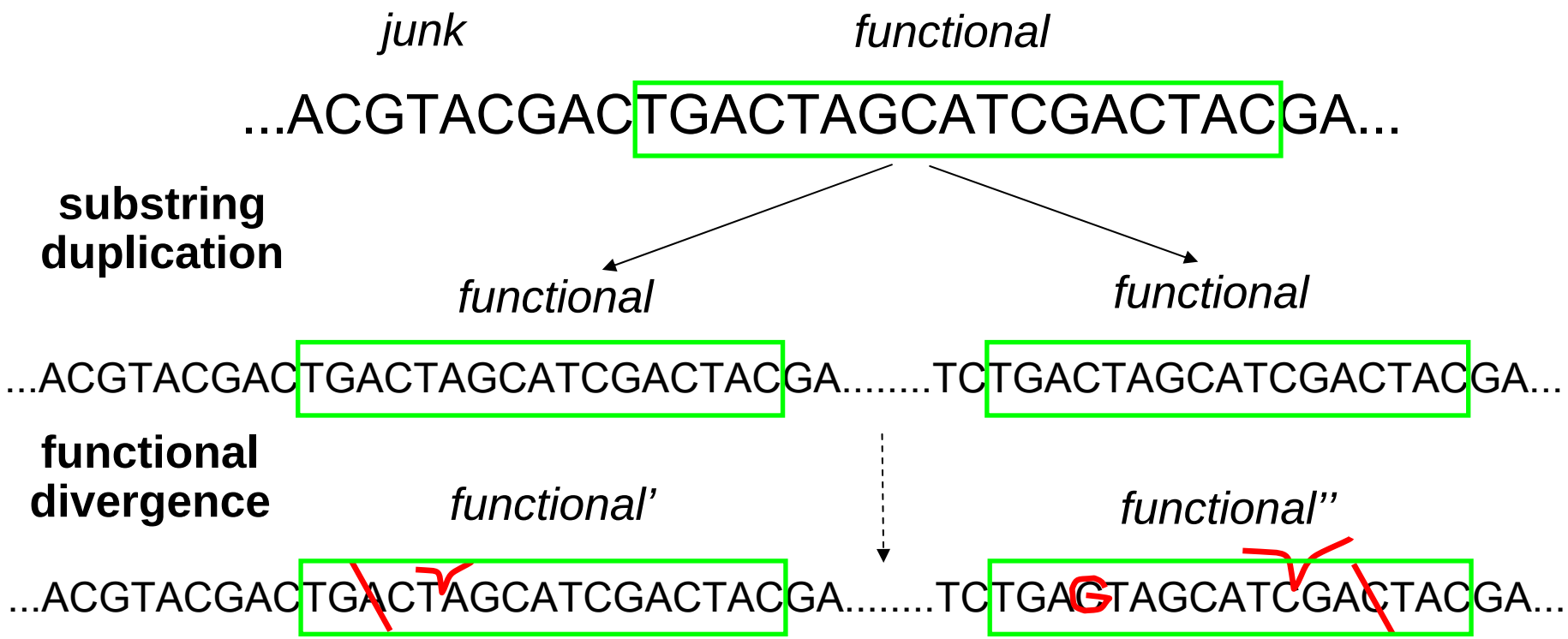
cell

"the workhorses of every living cell"

DNA Replication is Imperfect (contd)

Medium Scale: substrings are duplicated, deleted, inverted

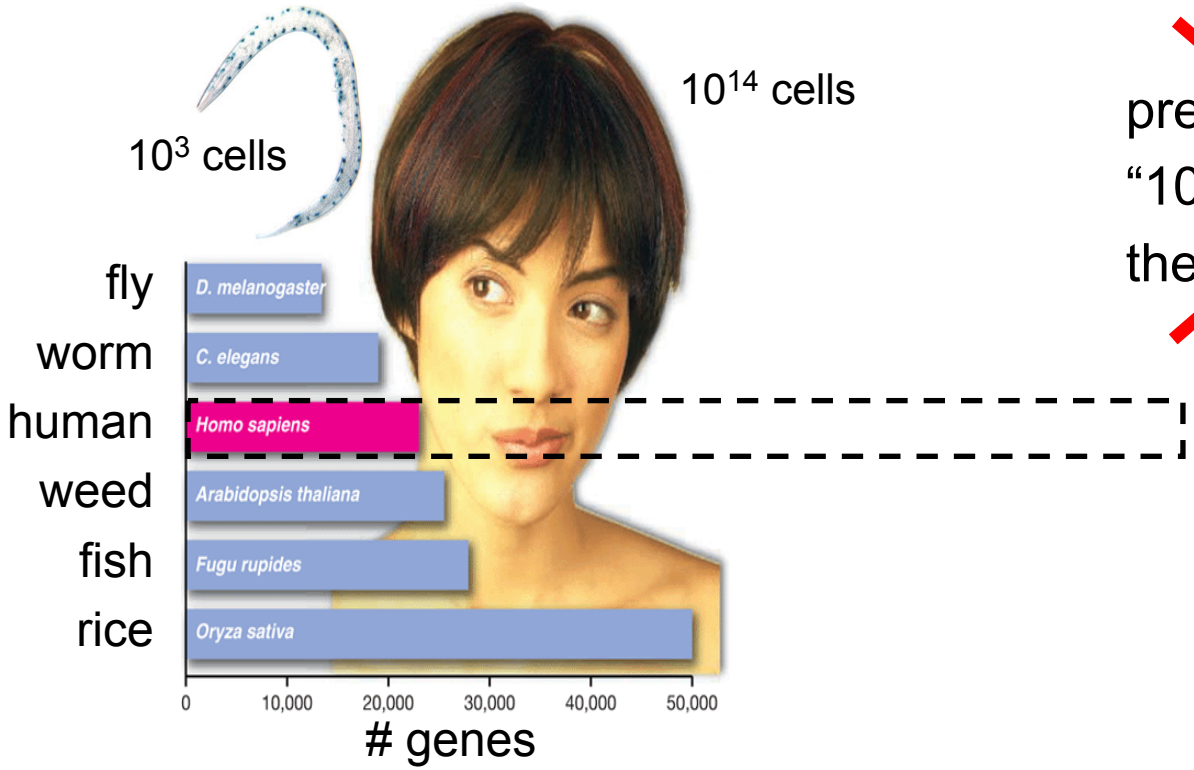
Large Scale: whole DNA strings are duplicated, deleted



So...More Genes...More Complexity!...Right?

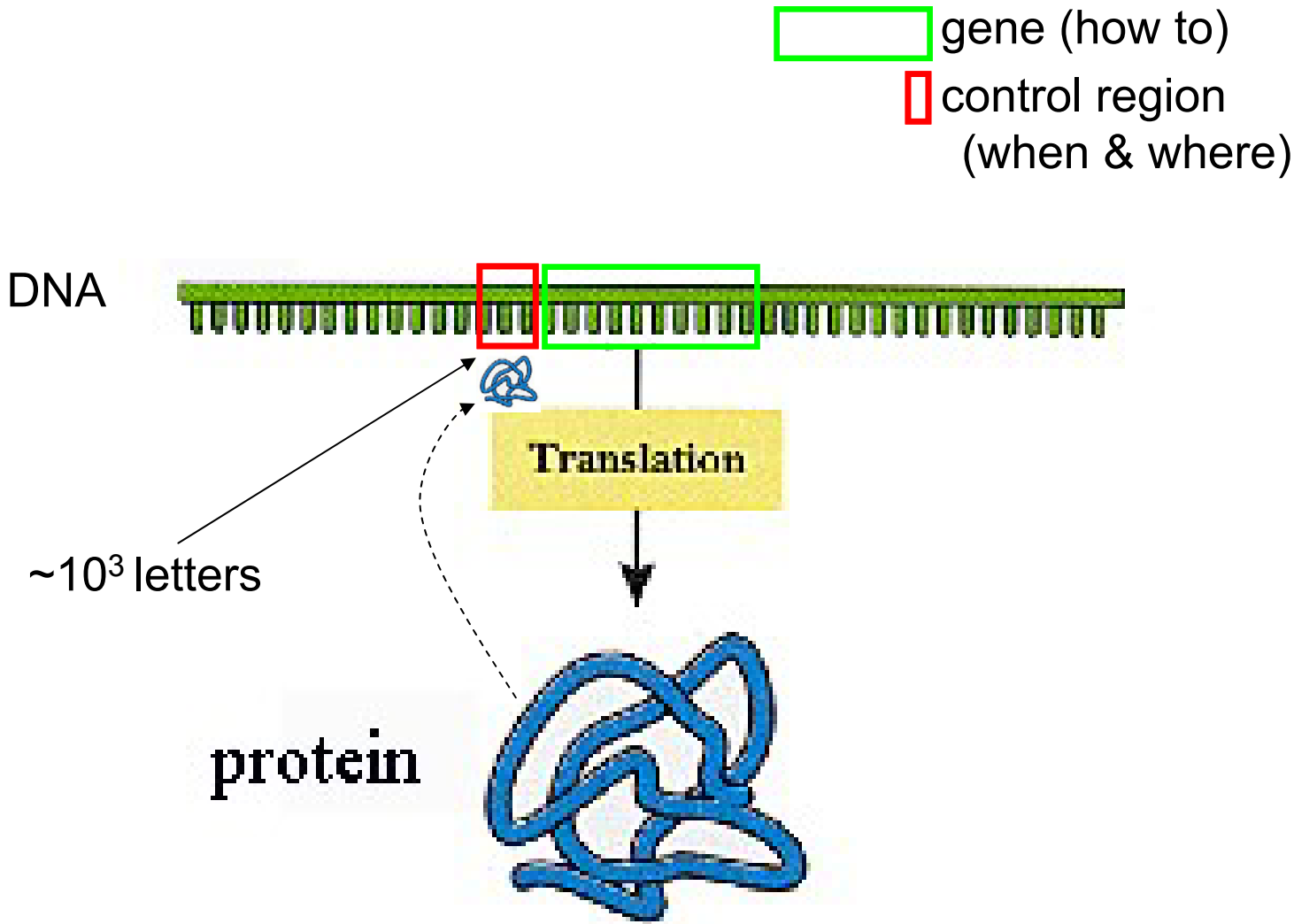
Genes & Complexity

Gene numbers do **not** correlate with organism complexity.
Many gene families are surprisingly old.

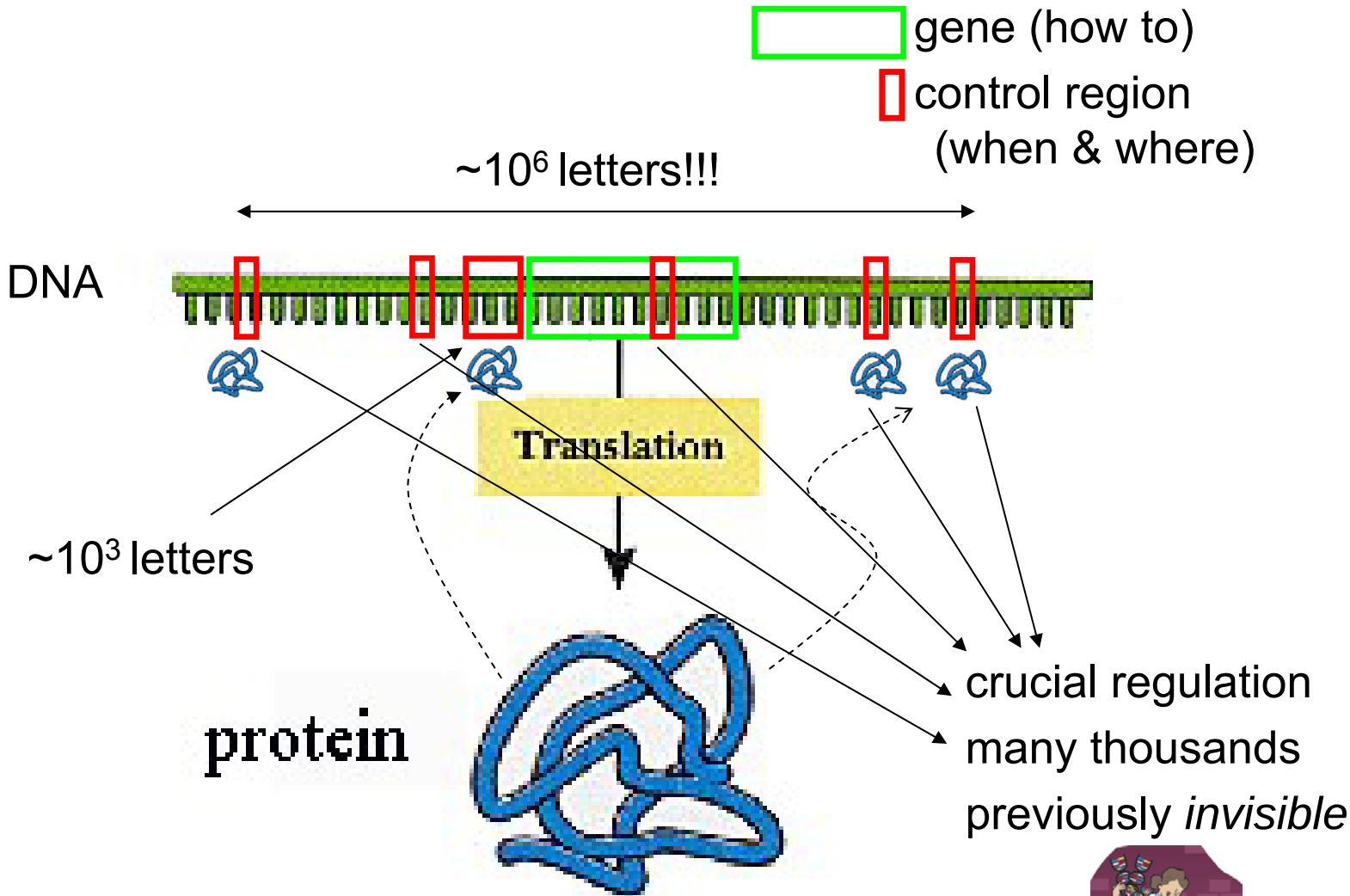


~~pre-genomic era:~~
~~"100,000 genes to the human genome"~~

Gene regulation = when/where to make protein

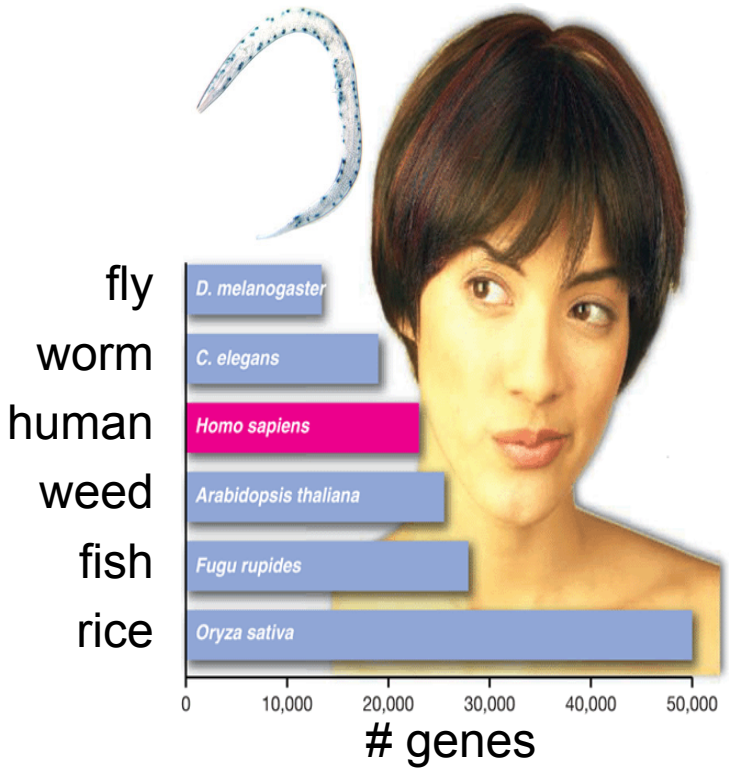


Vertebrate Gene Regulation

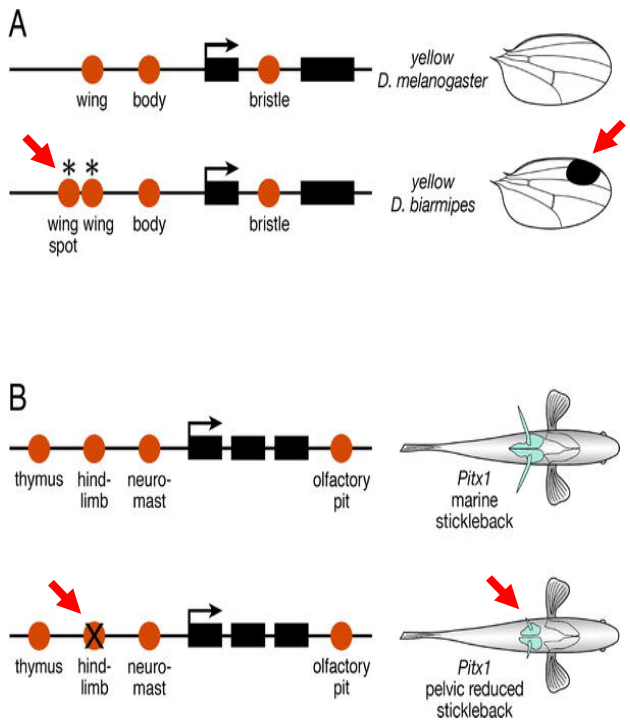


Regulatory regions drive morphological diversity

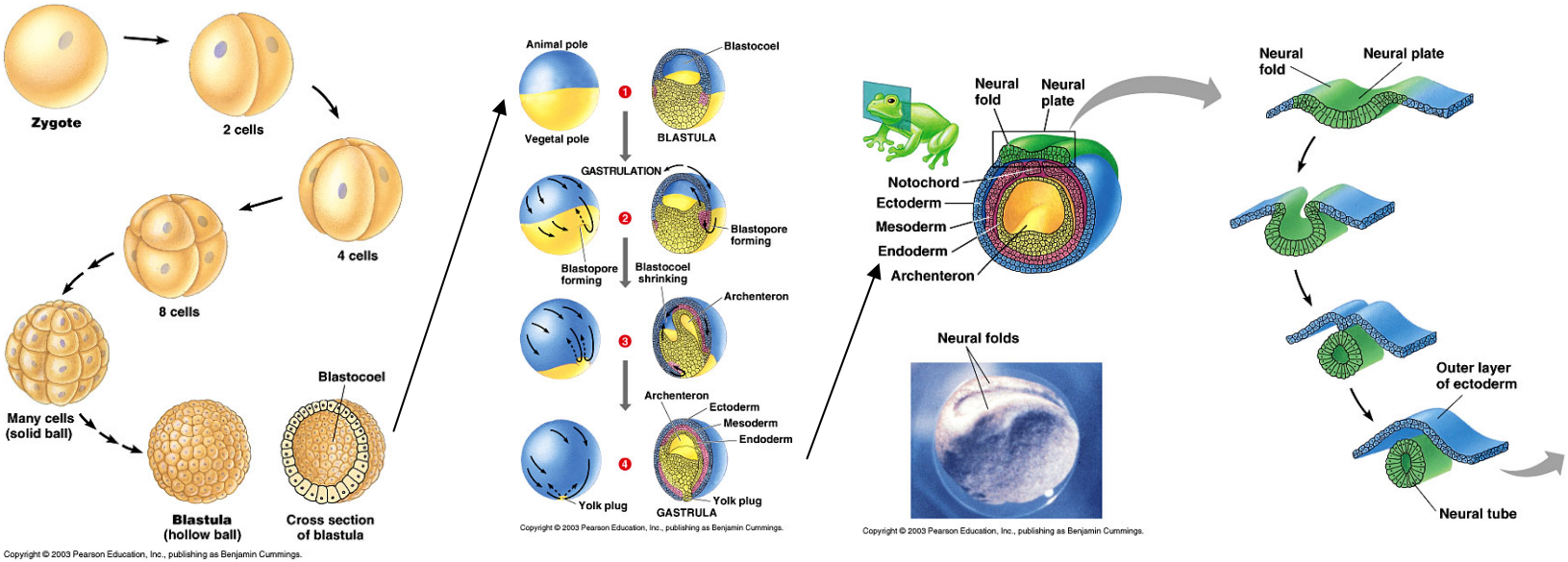
Gene numbers do **not** correlate with organism complexity.
Many gene families are surprisingly old.



"Regulatory sequence evolution must be the major contribution to the evolution of form." [Carroll, Wilson memorial lecture, *PLoS Biol,* 2005]



Embryonic Development and the Genome



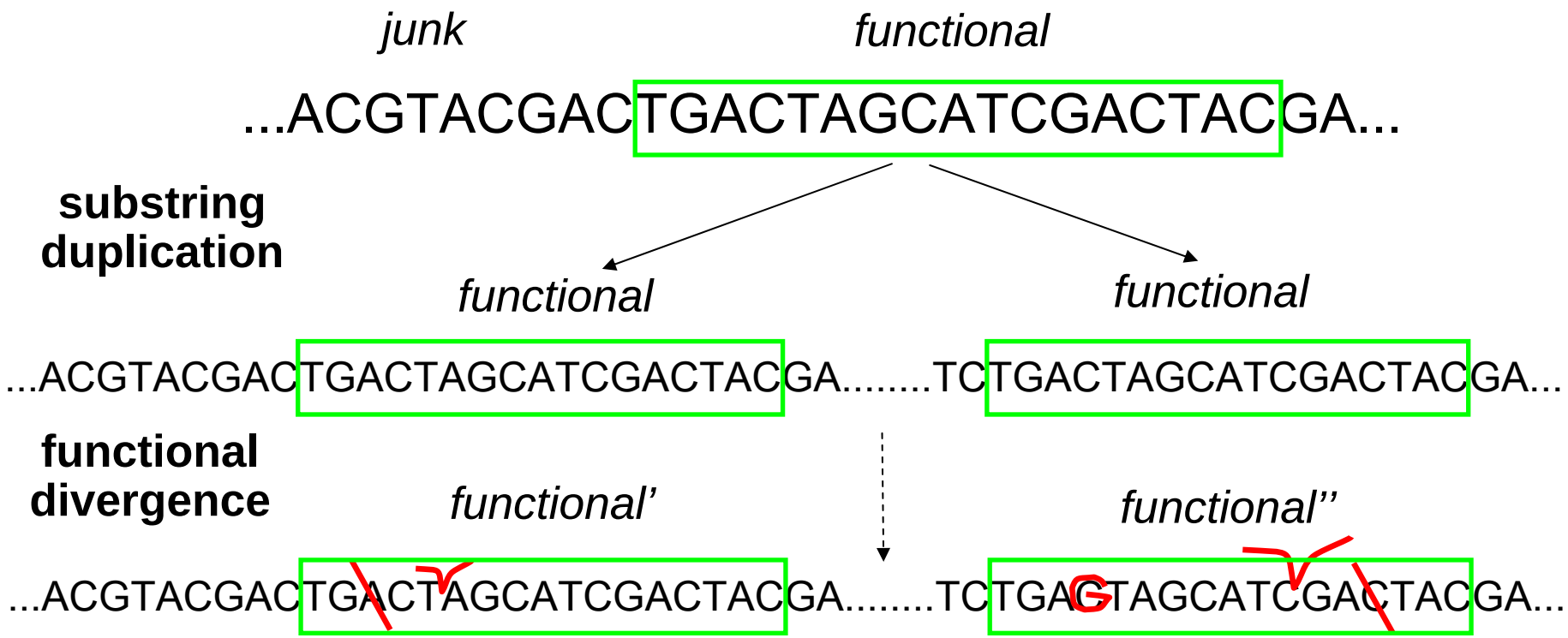
Many thousands of human conserved elements congregate en-masse near developmental genes.

[Eg, Dog Genome Paper, Nature, 2005; Bejerano et al., Nature Methods, 2005]

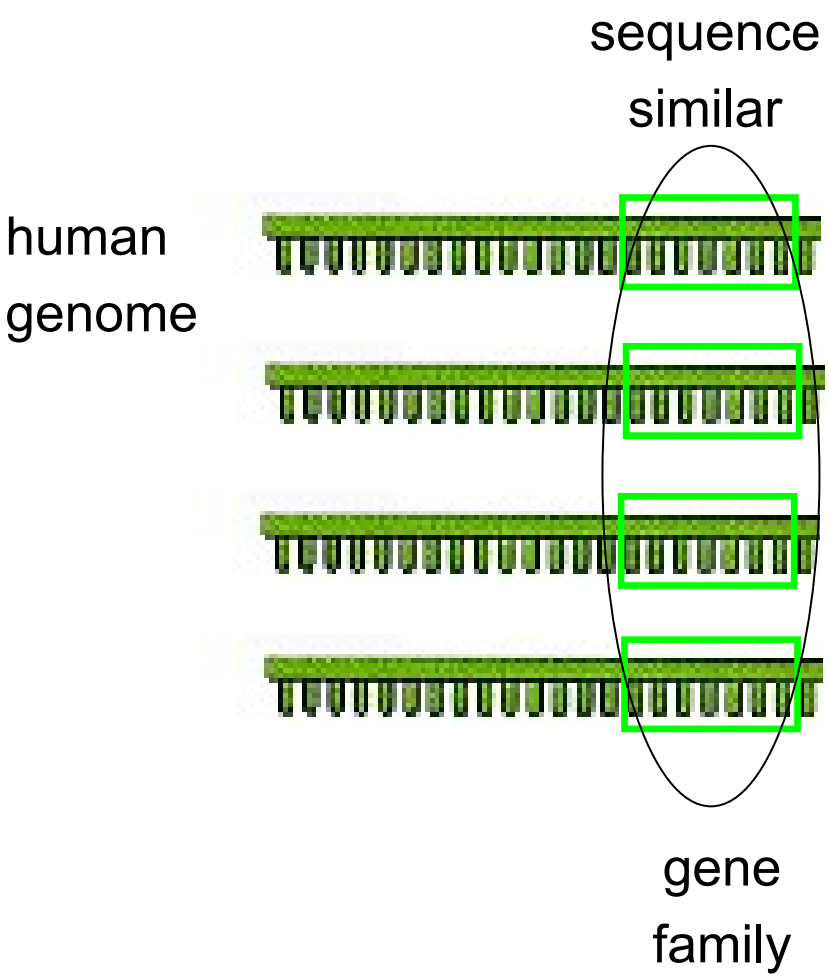
DNA Replication is Imperfect (reminder)

Medium Scale: substrings are duplicated, deleted, inverted

Large Scale: whole DNA strings are duplicated, deleted

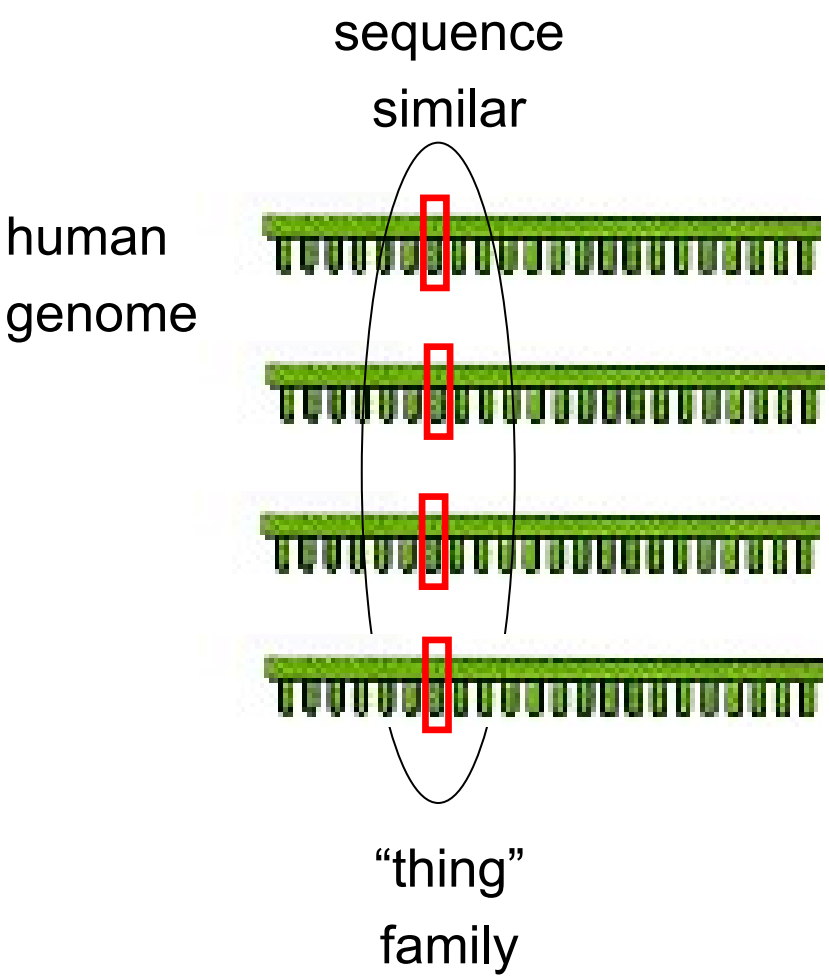


The Power of Computation



- similar function
- tale telling differences
- figure out one and you have a working hypothesis for all

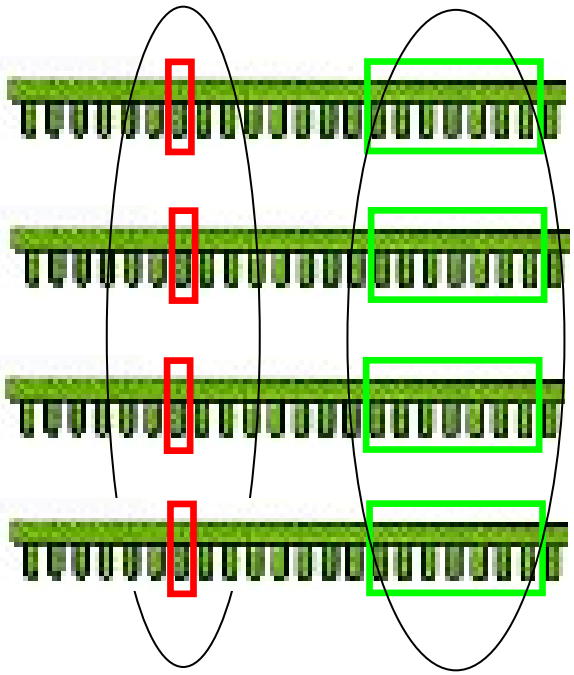
The Power of Computation



- similar function
- tale telling differences
- figure out one and you have a working hypothesis for all

The Power of Computation

human genome



"thing" family

gene family

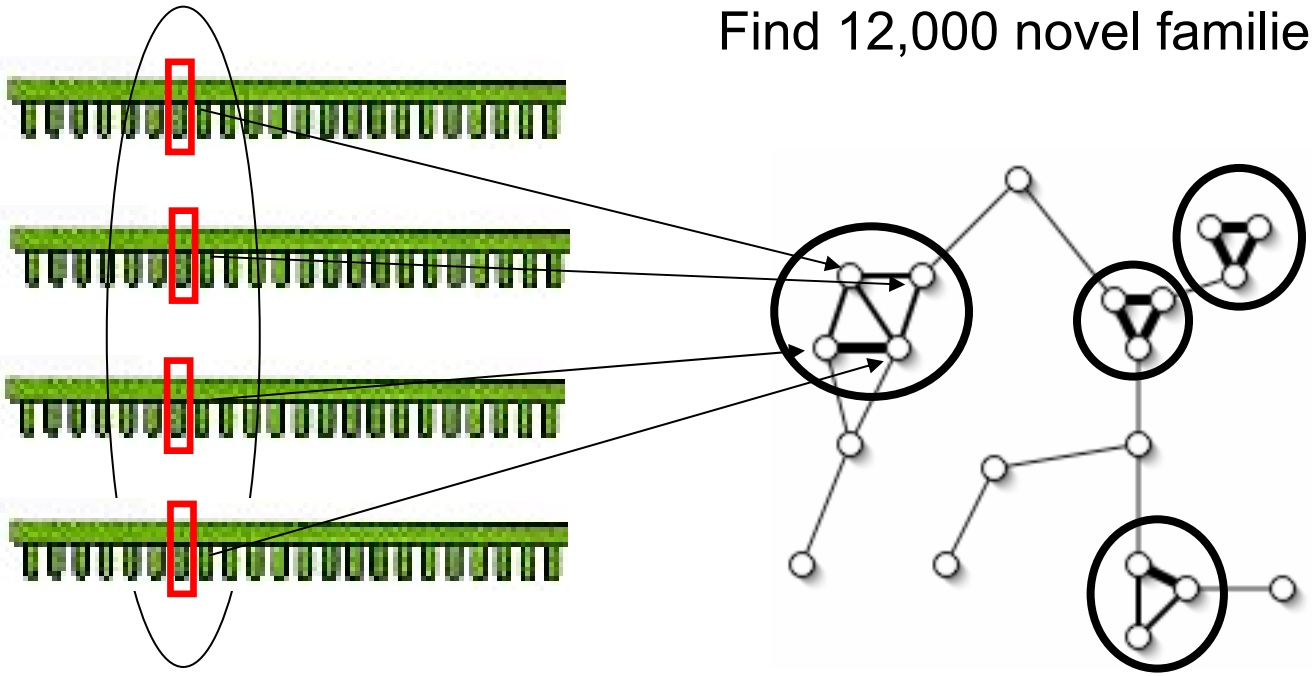
- similar function
- tale telling differences
- figure out one and you have a working hypothesis for all
- *"guilt by association" anchoring to genes (annotated landmarks)*

Families of Conserved Non Coding Elements

Cast as clustering problem

Find 12,000 novel families

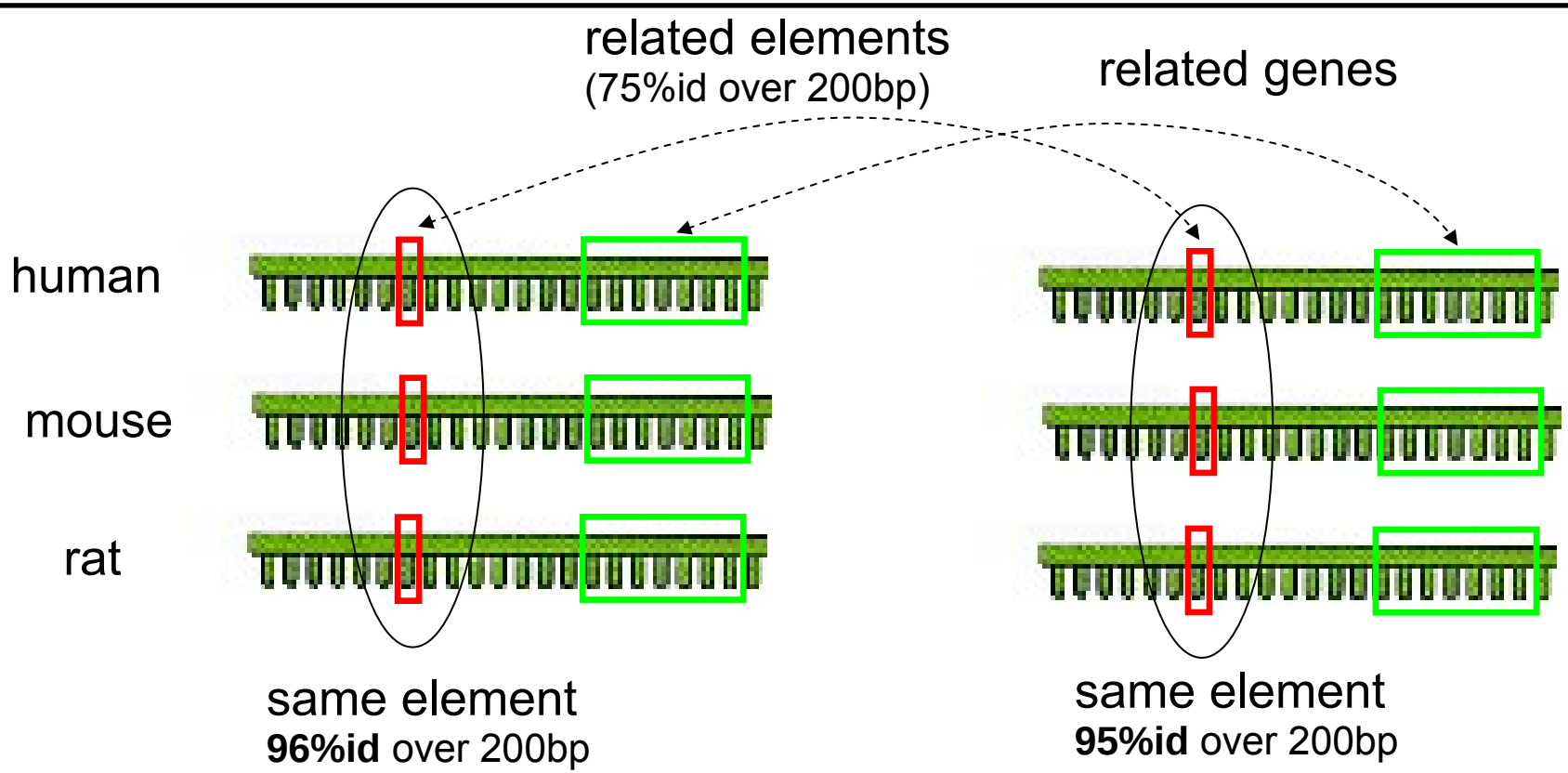
human genome



"thing" family

[Bejerano et al., *ISMB* 2004]

A Computational Question



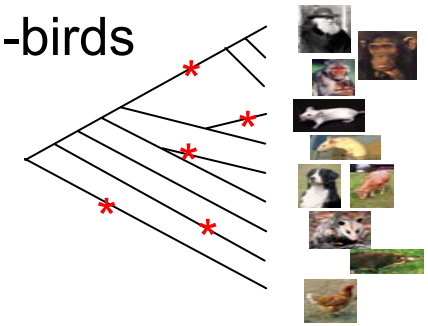
Classical Biological approach: experiment to understand *these* regions

Computational approach: how many regions *like* this are there?

Ultraconserved Elements

Hundreds of long substrings ***identical*** between human-birds
➔ they must have **rejected many different changes**.

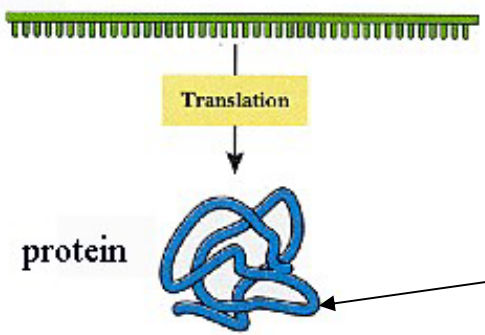
But... ***all*** functions we understand in our genome are encoded using **redundant codes**.



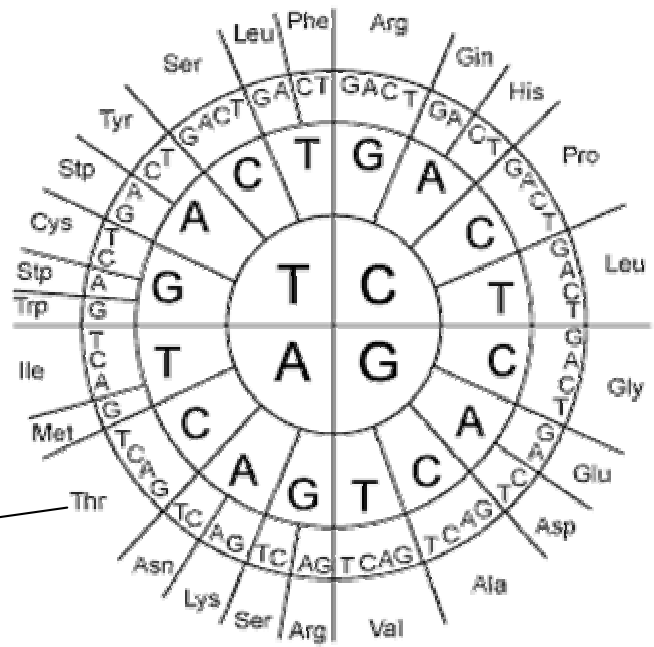
E.g. Protein Coding Genes:

DNA – 10^8 letters
over alphabet of 4.

Protein – 10^2 letters
over alphabet of 20.



Coding: 3 DNA letters → 1 Protein letter.

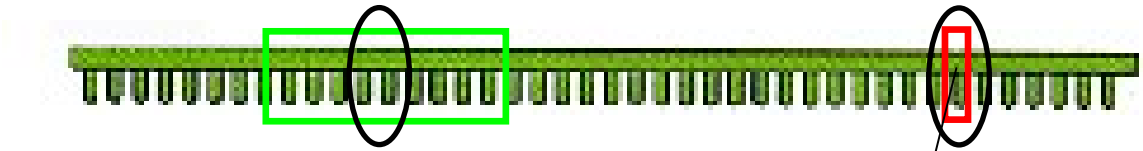


[Bejerano et al., *Science* 2004]

Computational Hypotheses

Based on public domain genome wide data:

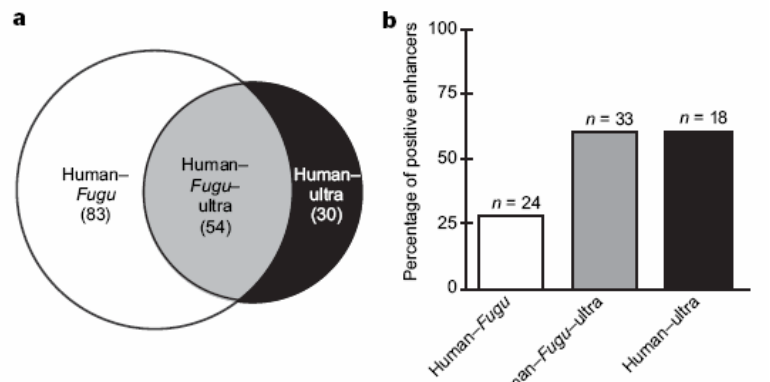
ultraconserved elements

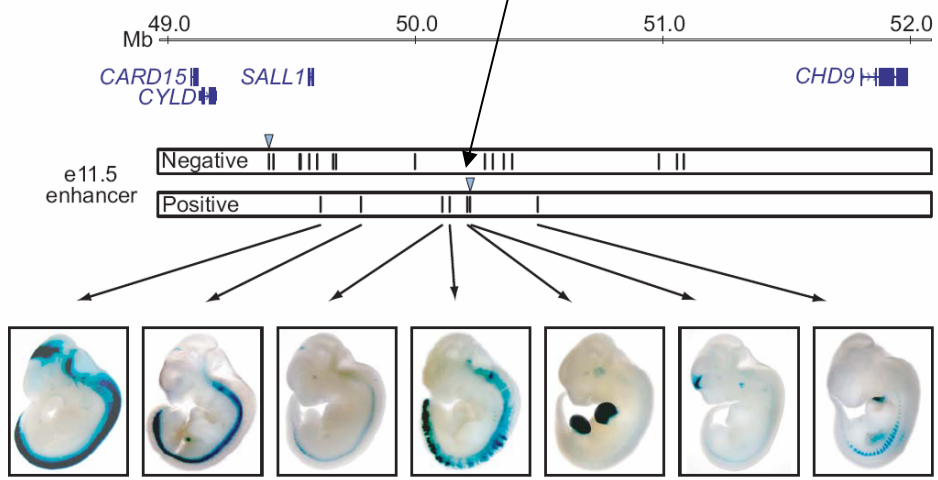


one subset codes protein

larger subset does not

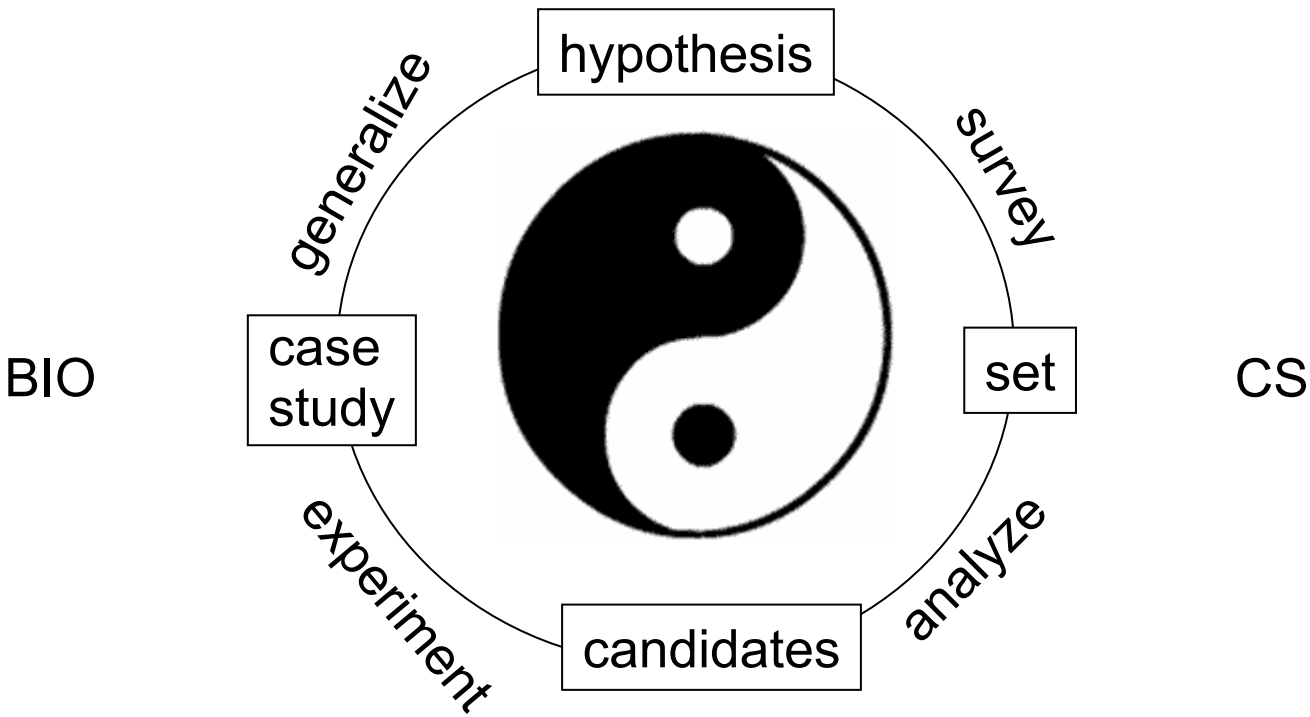
generate testable hypotheses for function from existing knowledge (2004)



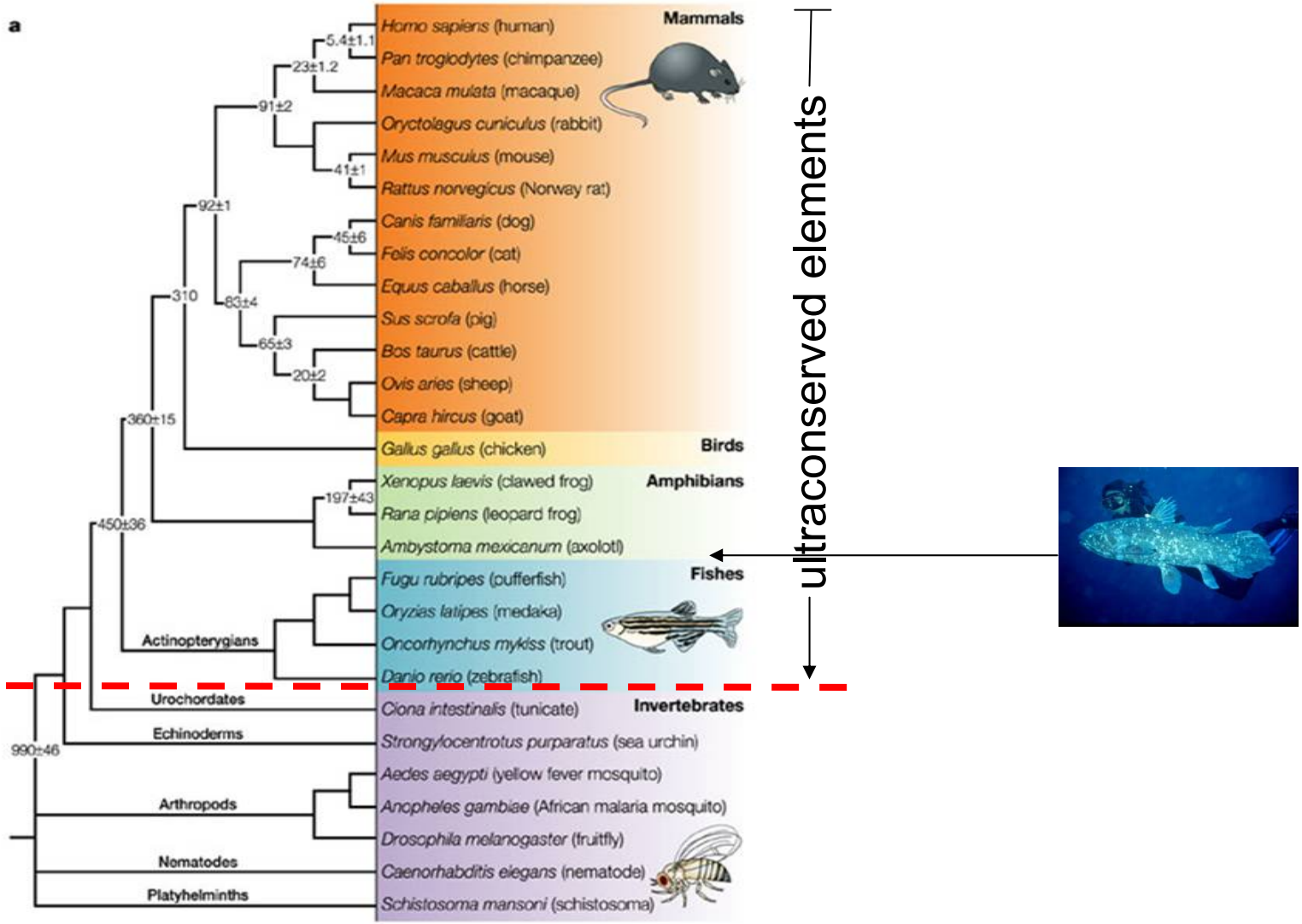


[Pennacchio et al., *Nature*, 2006]

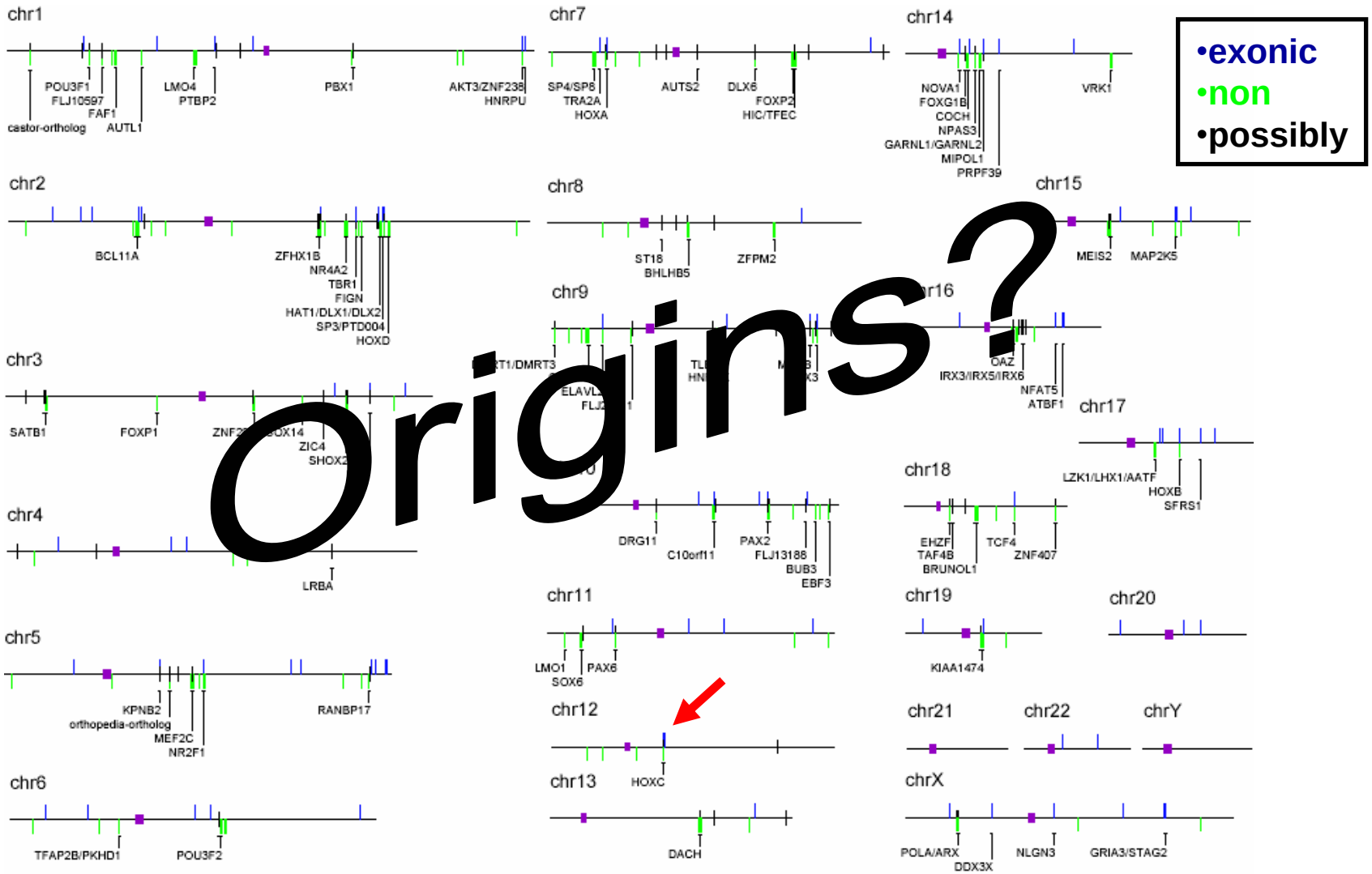
Computationally Driven Biology Simplified



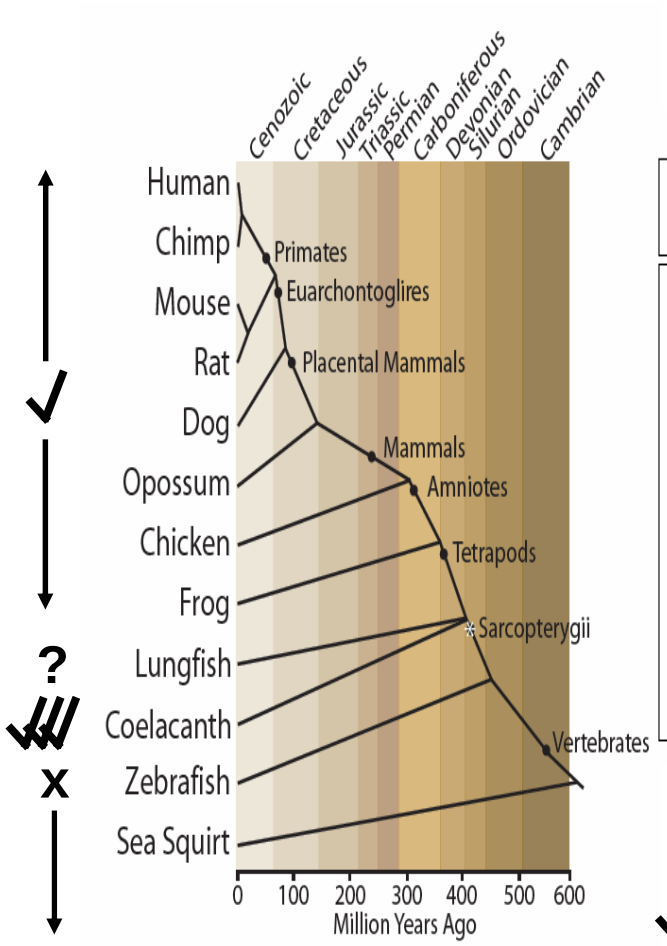
Origins of Ultraconserved Elements?



Genomic Distribution of Ultraconserved Elements



Uniquely Abundant in Coelacanth



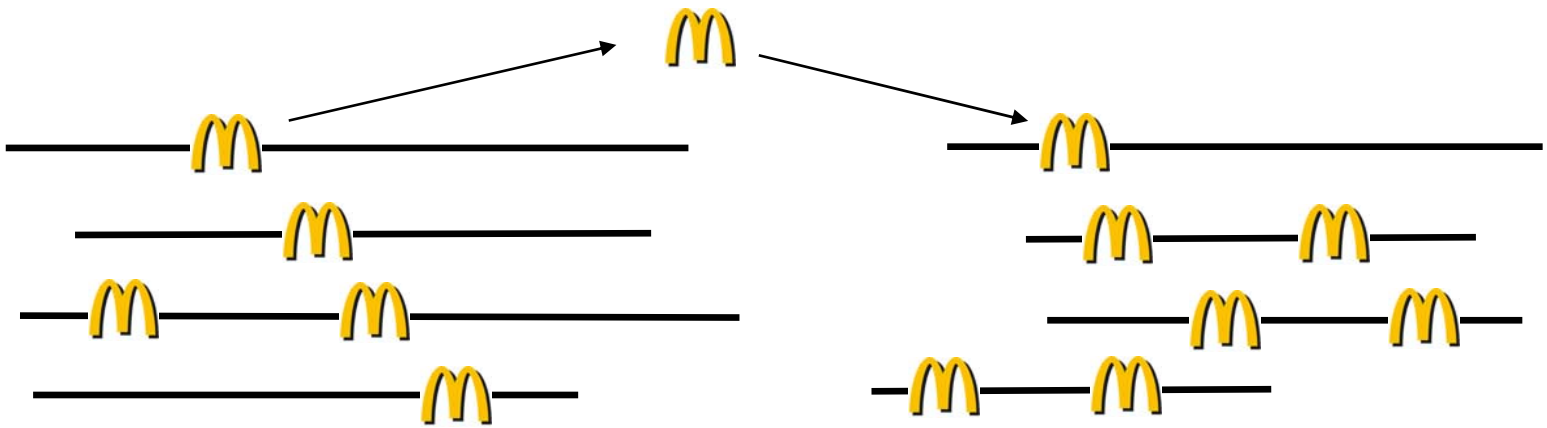
Upto 80%id between Coelacanth instances and some human instances, inc uc.338.

Species	UCSC Assembly	LF-SINE Detected	Species	UCSC Assembly	LF-SINE Detected
Homo sapiens	hg17	Yes	*Danio rerio*	danRer2	No
Pan troglodytes	panTro1	Yes	*Tetraodon nigroviridis*	tetNig1	No
Macaca mulatta	rheMac1	Yes	*Takifugu rubripes*	fr1	No
Mus musculus	mm6	Yes	*Ciona intestinalias*	ci1	No
Rattus norvegicus	rn3	Yes	*Strongylocentrotus purpuratus*	strPur1	No
Canis familiaris	canFam1	Yes	*Drosophila melanogaster*	dm2	No
Bos taurus	bosTau1	Yes	*Anopheles gambiae*	anoGam1	No
Monodelphis demestica	monDom1	Yes	*Caenorhabditis elegans*	ce2	No
Gallus gallus	galGal2	Yes	*Saccharomyces cerevisiae*	sacCer1	No
Xenopus tropicalis	xenTro1	Yes			

✓ 100 diverged copies in a Gigabase

✓✓✓ 60 highly similar copies in a Megabase

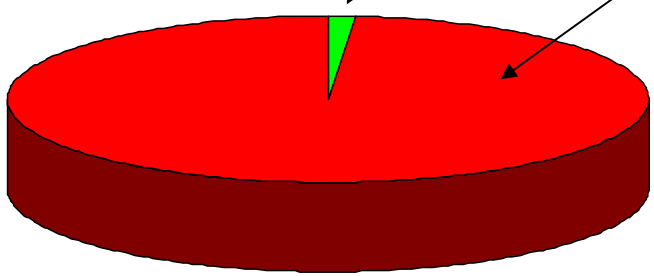
Repeats / Mobile Elements ("selfish DNA")



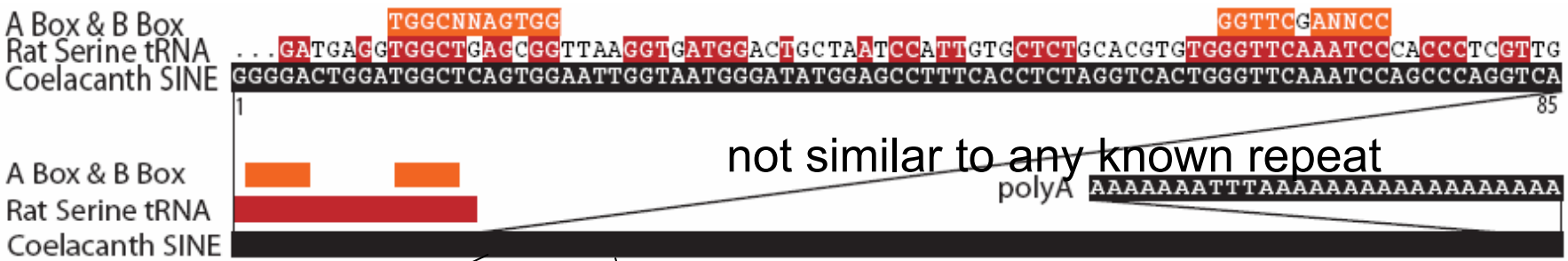
Human Genome: $3*10^9$ letters

1.5% known function

>50% junk



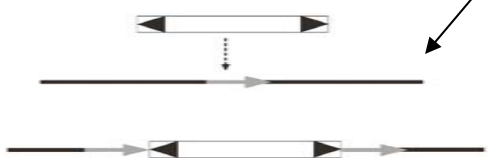
The LF SINE (for Lobefin Fish / “Living Fossil”)

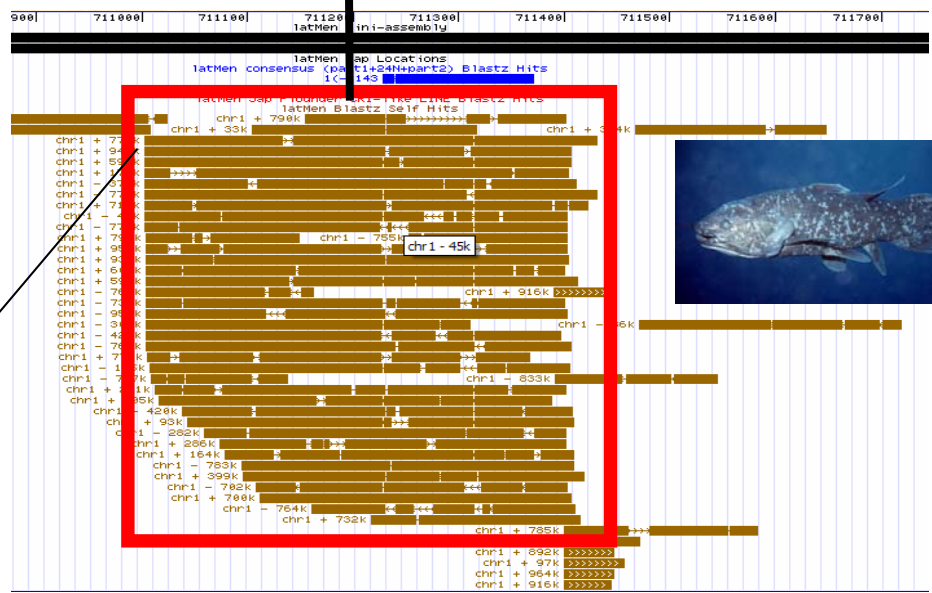


Reconstruction

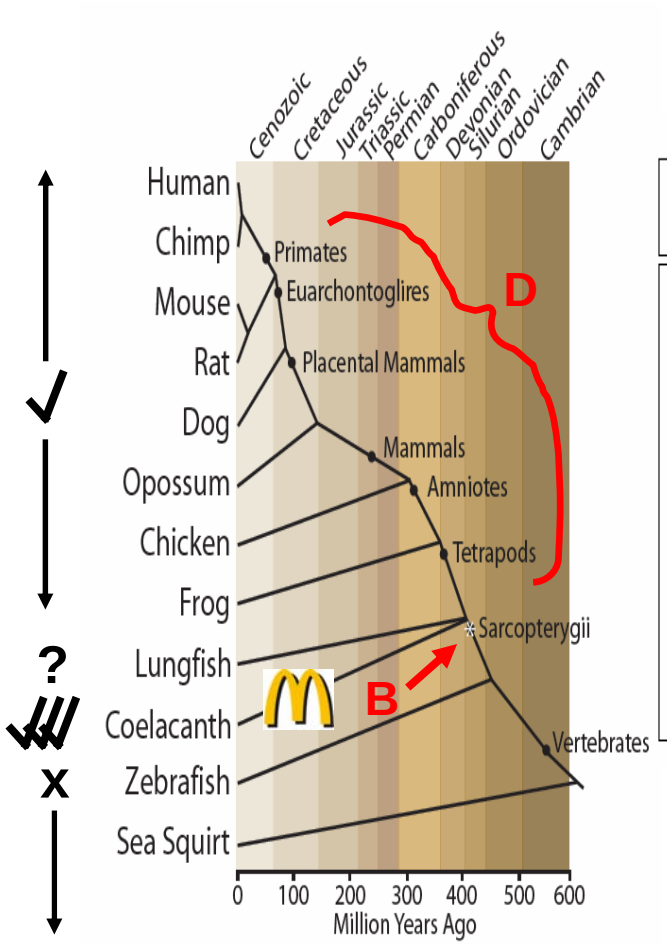
```
                100        110        120        130        140
LF-SINE         GAAAGTCA-TTACCATCTGATGGCTGTTCAGTGGCCTA---TGTGAAATGAG
                : :::::: ::: : :::    ::    :::::  :::    ::: ::::: :
Core Domain     GCAAGTCACTTAACCTCTCTGAGCC--TCAGTTTCCTCATCTGTAAAATGGG
                  100        110        120        130        140

                          90         100            110        120
V-SINE(shark)   CTGTGTGGAGTTTGCACATTCTC----CCCGTGTCTGCGTGGGTTTCCTC
                :::::::    :: :   : :: :    ::   : :  :: :  ::::::::
Core Domain     CTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGAGCCTCAGTTTCCTC
                          90         100        110        120        130
```

target site duplications





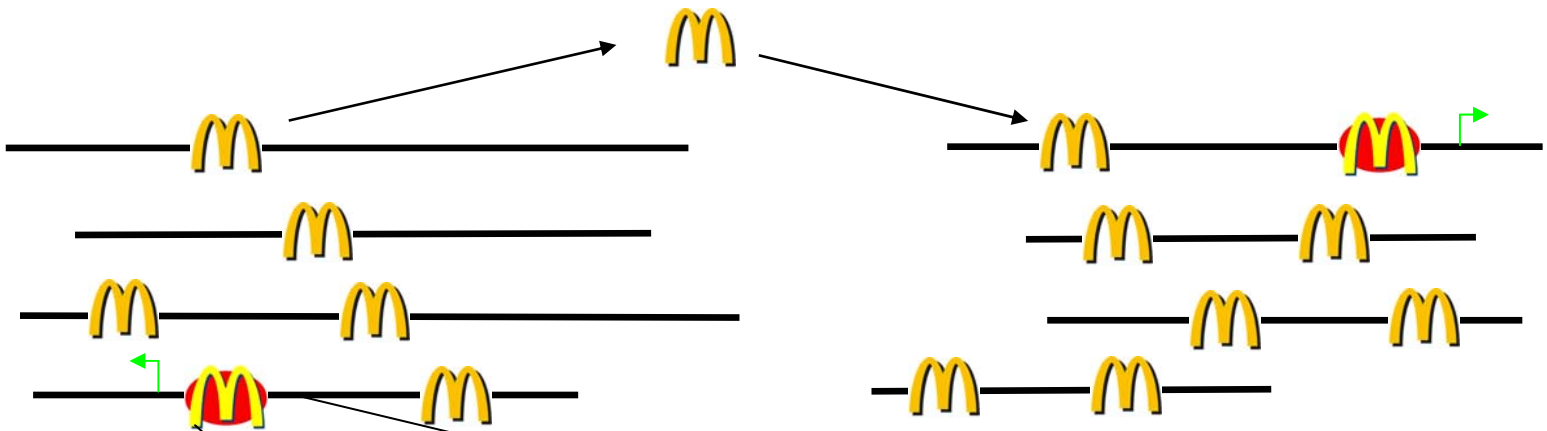
>360My Old and Going Strong



Upto 80%id between Coelacanth SINE and some human instances, inc uc.338.

Species	UCSC Assembly	LF-SINE Detected	Species	UCSC Assembly	LF-SINE Detected
Homo sapiens	hg17	Yes	*Danio rerio*	danRer2	No
Pan troglodytes	panTro1	Yes	*Tetraodon nigroviridis*	tetNig1	No
Macaca mulatta	rheMac1	Yes	*Takifugu rubripes*	fr1	No
Mus musculus	mm6	Yes	*Ciona intestinalias*	ci1	No
Rattus norvegicus	rn3	Yes	*Strongylocentrotus purpuratus*	strPur1	No
Canis familiaris	canFam1	Yes	*Drosophila melanogaster*	dm2	No
Bos taurus	bosTau1	Yes	*Anopheles gambiae*	anoGam1	No
Monodelphis demestica	monDom1	Yes	*Caenorhabditis elegans*	ce2	No
Gallus gallus	galGal2	Yes	*Saccharomyces cerevisiae*	sacCer1	No
Xenopus tropicalis	xenTro1	Yes			

Cis-reg & Ultra elements from Mobile Elements



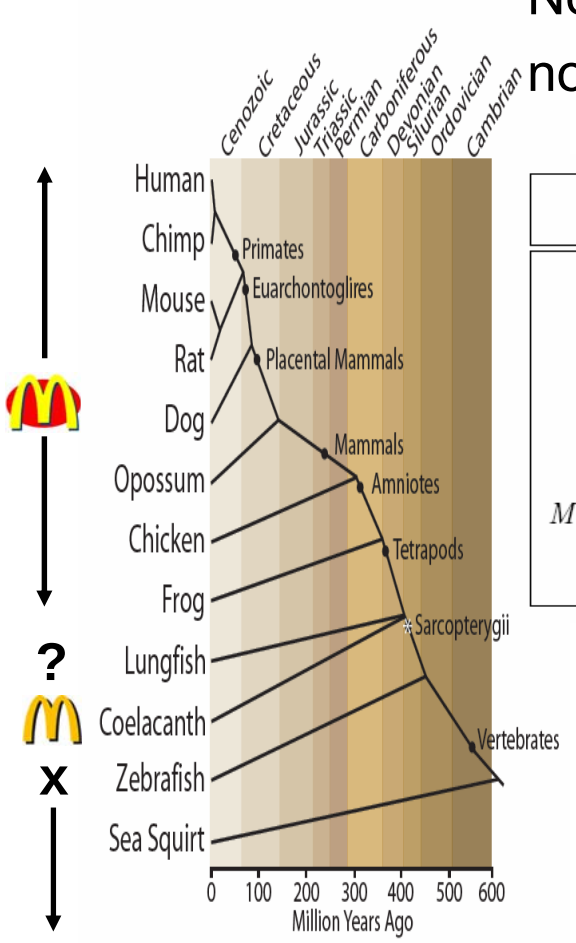
Co-option event, probably due to favorable genomic *context*

All other copies are destined to decay over time at a neutral rate

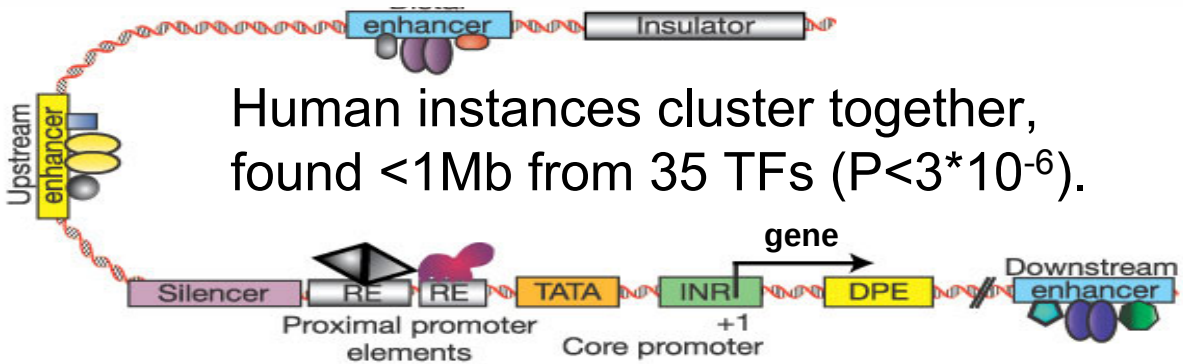
[Yass is a small town in New South Wales, Australia.]

Exapted Into Which Cellular Roles?

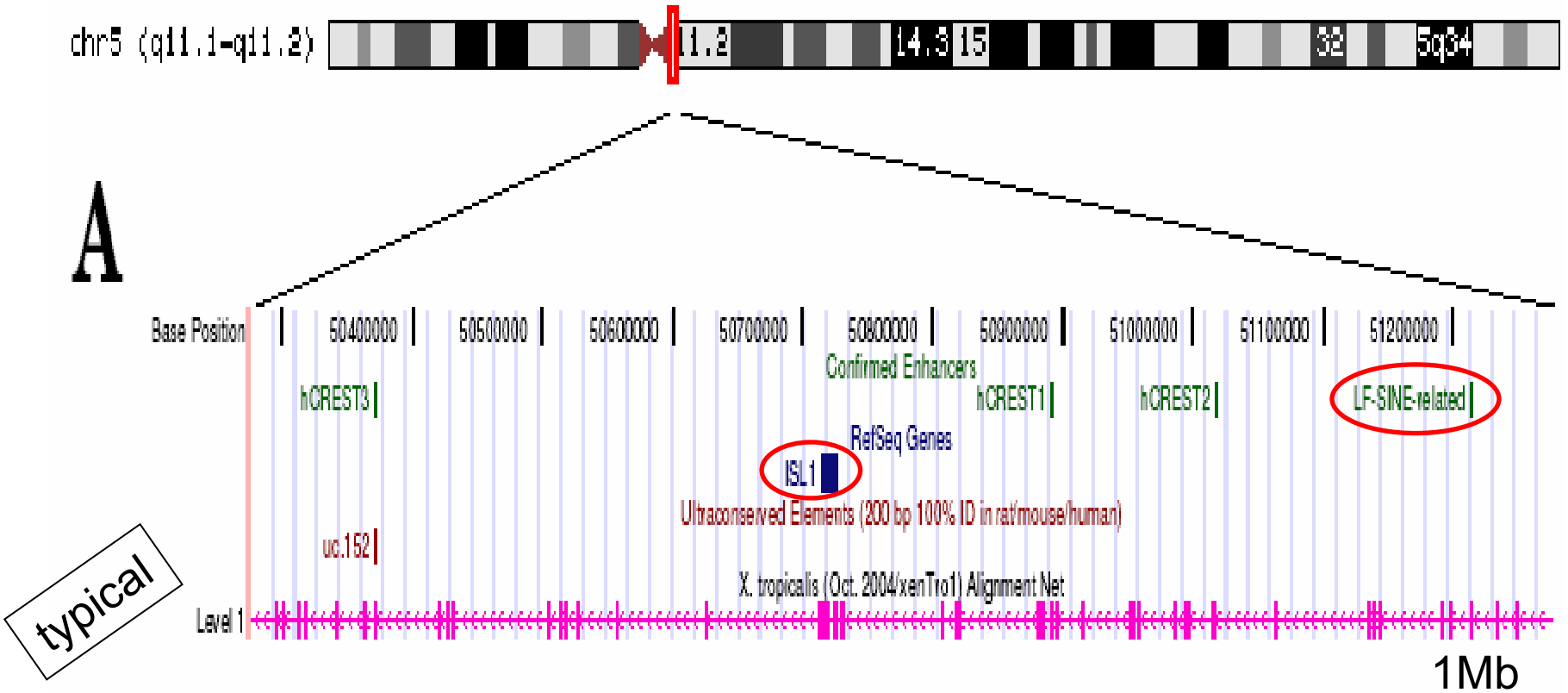
No evidence for Transcription (Tx) as small RNAs,
no orientation preference in introns, not in antisense Tx.



Organism	5' UTR	3' UTR	Exonic Alt-Spliced	Exonic Total	Intronic	Intergenic	Total
Homo sapiens	1	0	12	13	68	163	245
Pan troglodytes	-	-	-	-	-	-	210
Macaca mulatta	-	-	-	-	-	-	229
Canis familiaris	-	-	-	-	-	-	235
Bos taurus	-	-	-	-	-	-	169
Mus musculus	0	1	7	8	25	57	91
Rattus norvegicus	-	-	-	-	-	-	87
Monodelphis domestica	-	-	-	-	-	-	394
Gallus gallus	0	1	2	3	244	451	699
Xenopus tropicalis	0	0	1	2	10	14	26



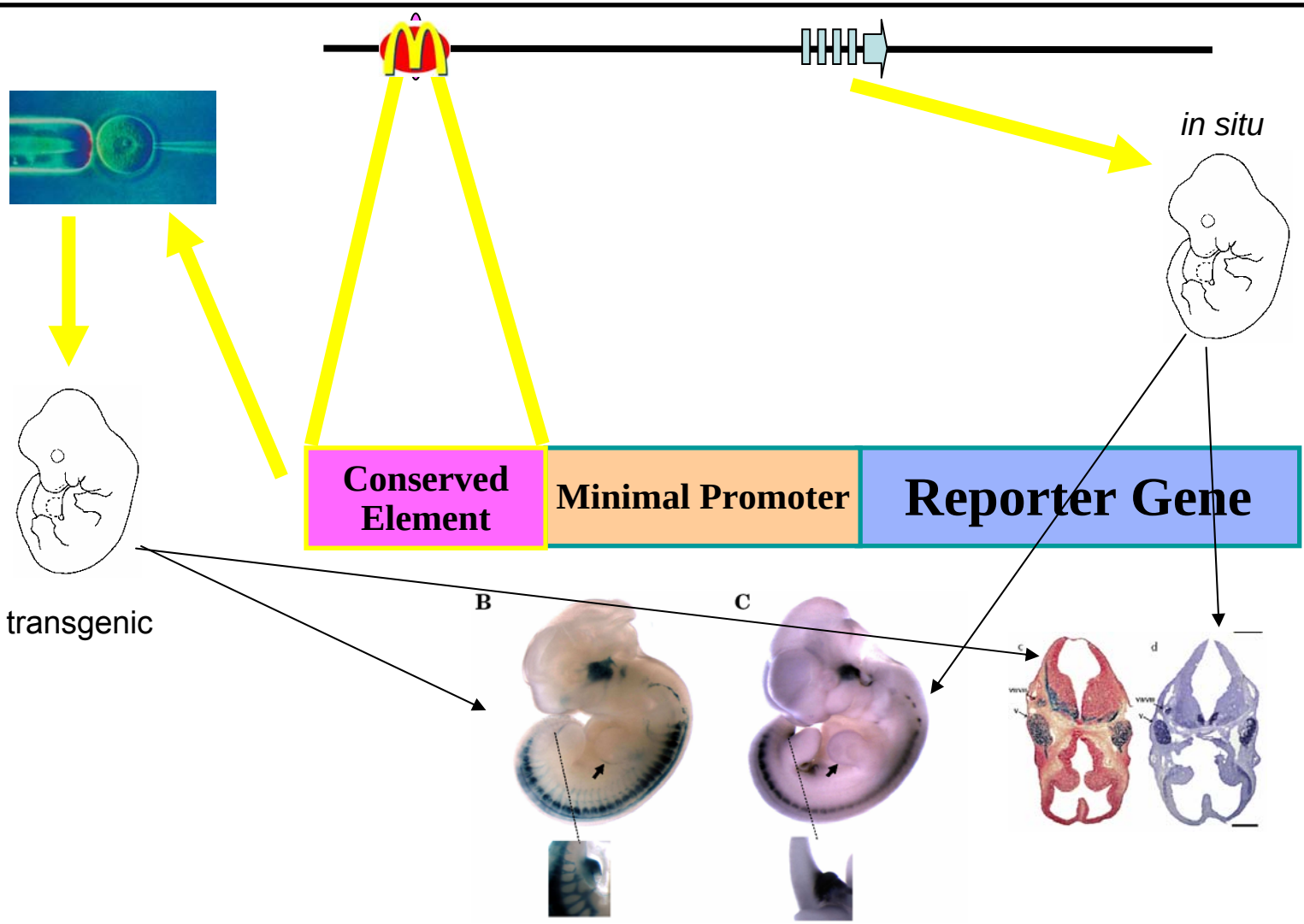
Instance 500kb Downstream of ISL1



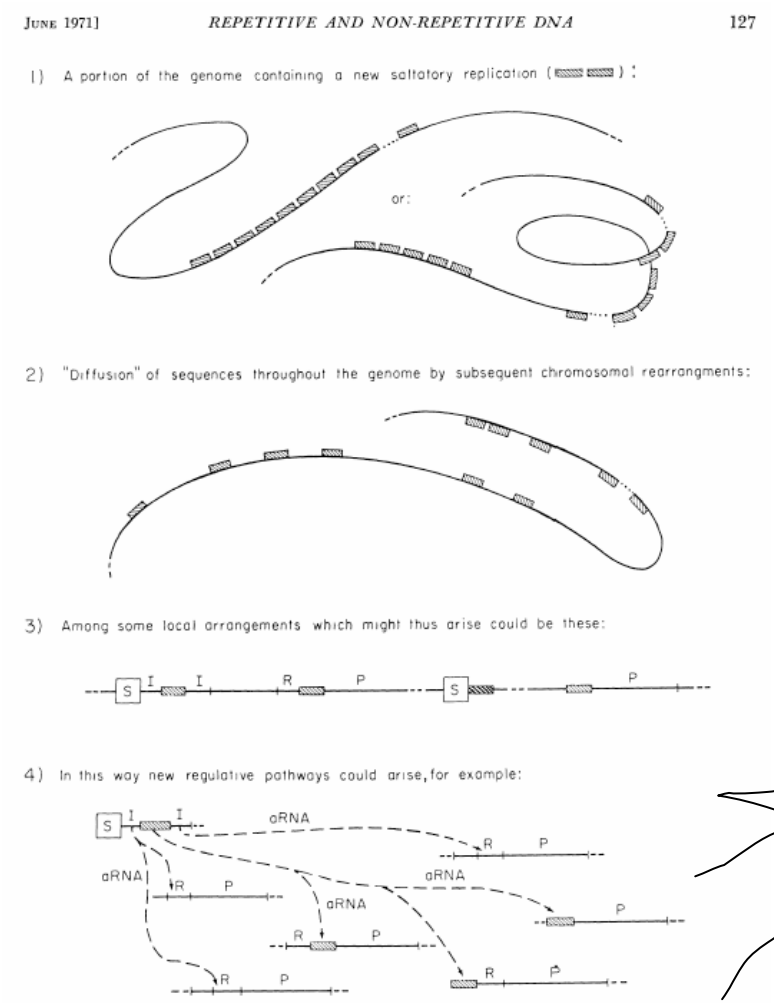
ISL1 is a neuro-developmental gene, also expressed in testis.

Three previously known enhancers are conserved across vertebrates.

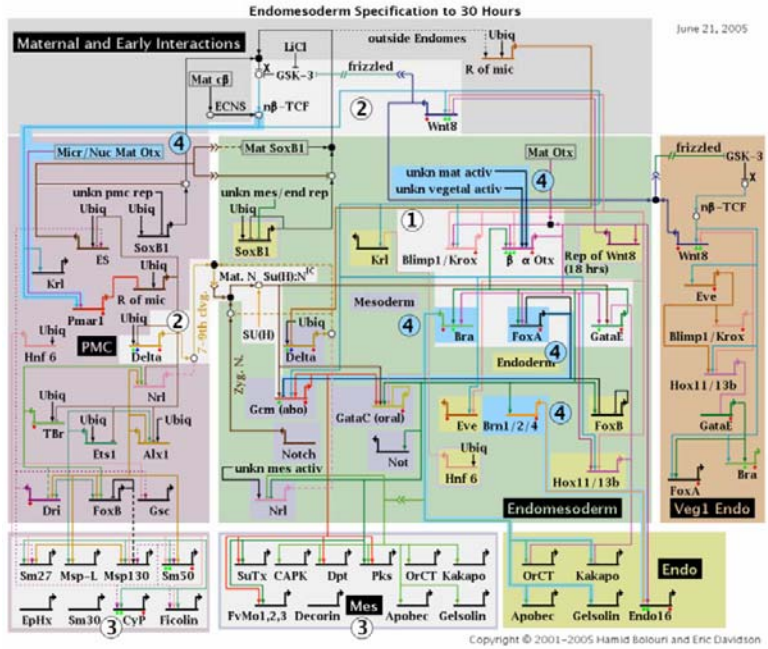
Repeat made Regulatory Region



From junk DNA to pathway recruitments?



[Britten & Davidson, 1971]

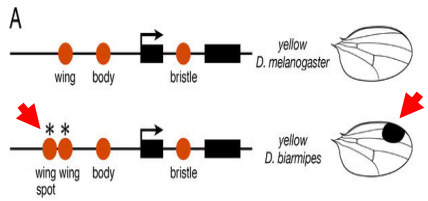


[Davidson & Erwin, 2006]

The Co-Optionome

quantify co-option

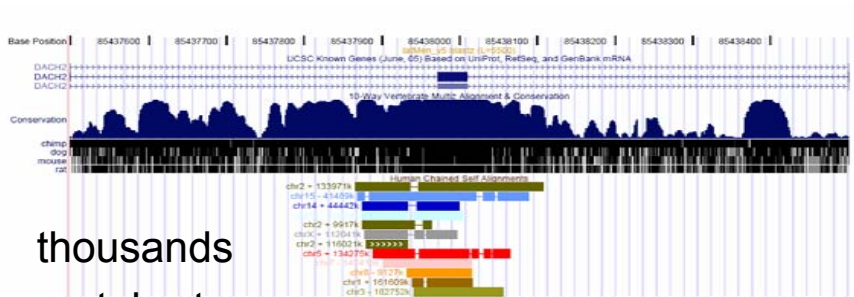
characterized repeats



functional elements

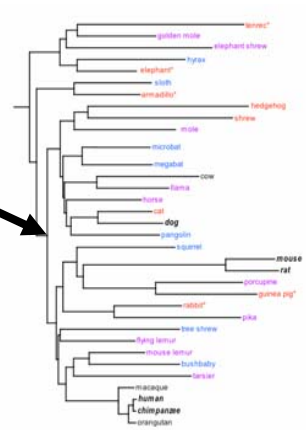
genomic DNA

repeat detectors



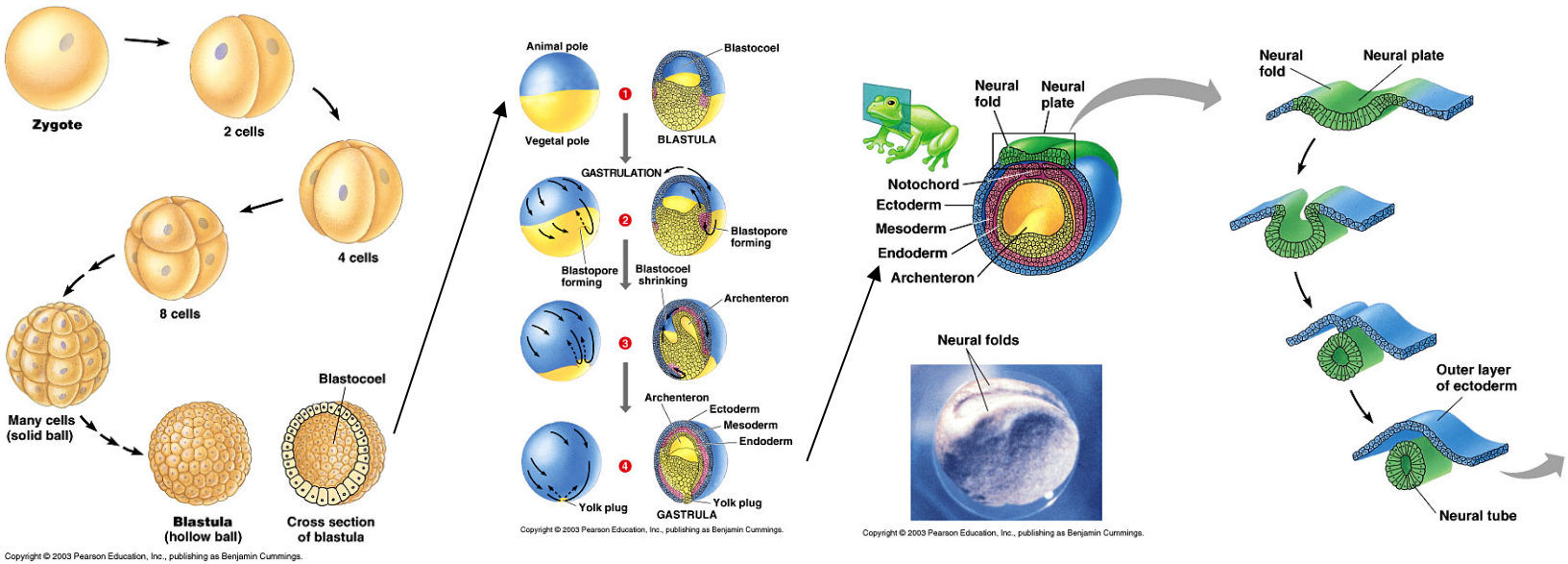
thousands vertebrate candidates

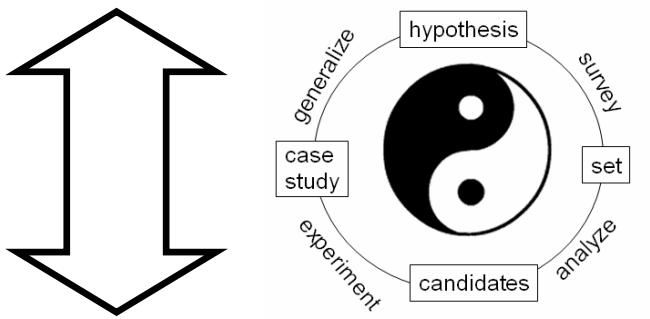
reconstructing Boreoeutherian ancestor

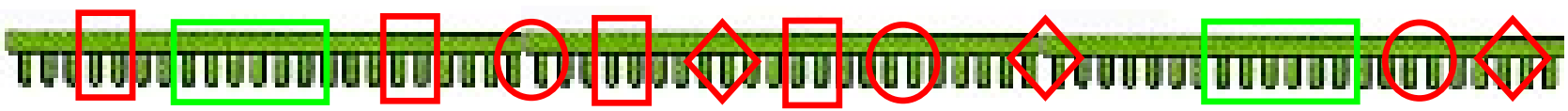


DeuSINE, MER121, …

Objective: Marry Development & Genomics





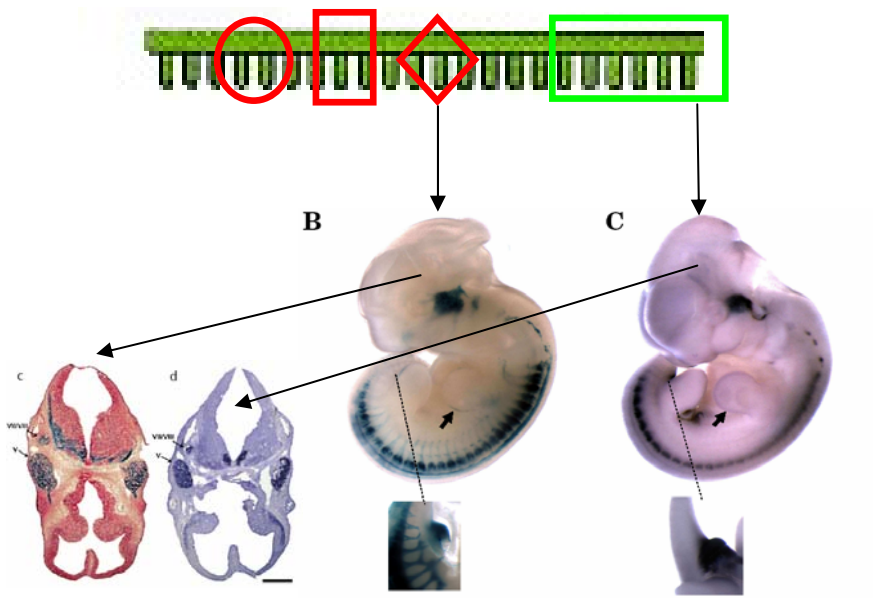


Bejerano Lab: Research Interests

Many thousands of human conserved elements congregate en-masse near developmental genes.

- Origins & Evolution
- Functions & Encoding
- Contribution to Human Disease

Break regulatory code



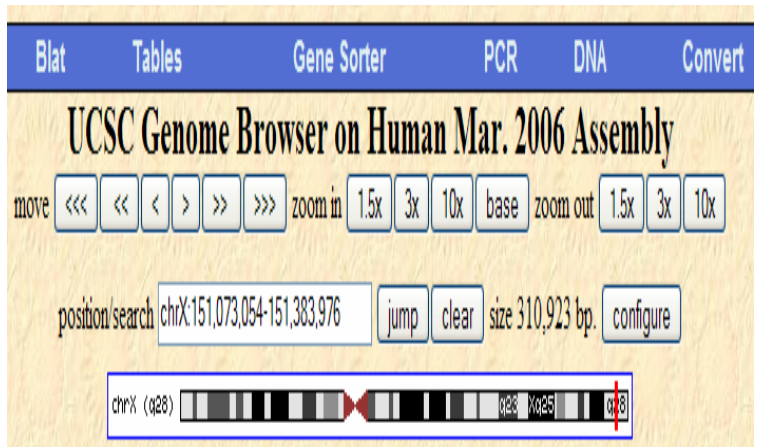
Bejerano Lab: Research Interests

Many thousands of human conserved elements congregate en-masse near developmental genes.

- Origins & Evolution
- Functions & Encoding
- Contribution to Human Disease

Break regulatory code

Provide tools to community



thousands and thousands of page requests served *daily*

Bejerano Lab: Research Interests

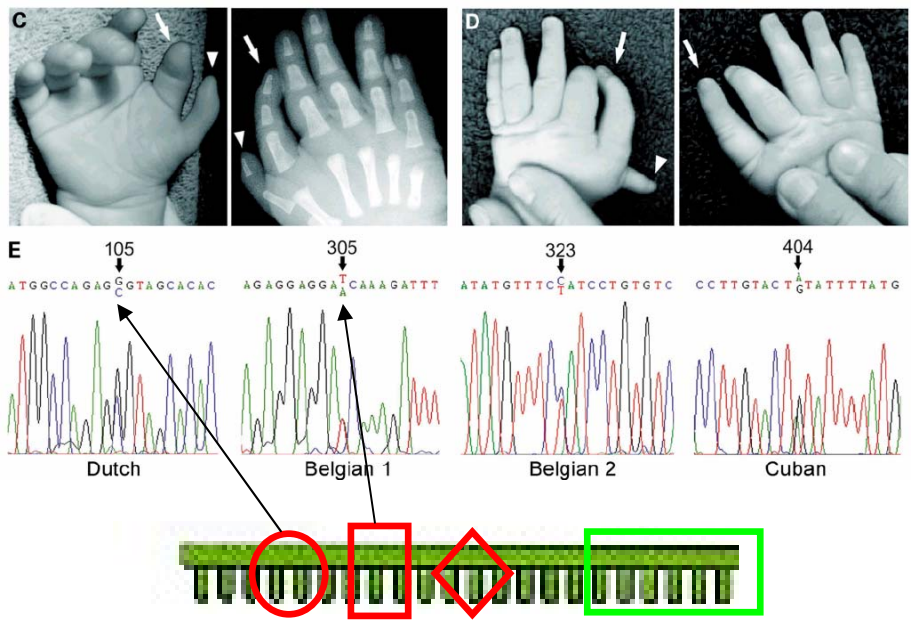
Many thousands of human conserved elements congregate en-masse near developmental genes.

- Origins & Evolution
- Functions & Encoding
- Contribution to Human Disease

Break regulatory code

Provide tools to community

Improve human health



Kudos

UC Santa Cruz

David Haussler

Sofie Salama, Jim Kent, Craig Lowe,

Bryan King, Adam Siepel, Jakob Pedersen

Katie Pollard, Courtney Onodera

Rachel Harte, Genomics/Browser Group

Lawrence Berkeley Labs

Eddy Rubin

Nadav Ahituv

McGill U.

Mathieu Blanchette

Penn State U.

Webb Miller's group

U. Queensland

John Mattick's group

Genome Sequencing Consortia

All GenBank contributors

Gill Bejerano

bejerano@stanford.edu